Amplitude Modulation of Noise in Voiced Fricatives: Acoustics, Psychoacoustics and Perception

Jonathan Pincas

Submitted for the Degree of
Doctor of Philosophy
from the
University of Surrey

Centre for Vision, Speech and Signal Processing
Faculty of Engineering and Physical Sciences
University of Surrey
Guildford, Surrey GU2 7XH, U.K.

June 2009

Summary

In voiced frication, the noise component exhibits amplitude modulation (AM). Psychoacoustic studies suggest that AM at the levels observed in voiced fricatives (VFs) is detectable and its inclusion in synthesis has improved quality, but it is not known whether listeners can exploit AM as a cue to the phonological voicing distinction in fricatives. A set of complementary acoustic, psychoacoustic and speech perception studies were conducted to test such a hypothesis.

Since little data was available on the acoustic properties of AM in VFs, a technique was developed to extract estimates of the depth of modulation, m, and applied to corpora of fricatives. Modulation depth estimates were compared to voicing strength, revealing a pattern for modulation at f_0 to rise at low voicing strengths and subsequently saturate. Saturation of m occurred at a voicing strength of 60–65 dB SPL depending on subject and experimental conditions. Modulation depths at saturation varied little across speakers but significantly for place of articulation; for example, modulation depth immediately after saturation was largest for [z] (0.65; cf. 0.44 for [ʒ], 0.37 for [ð], 0.34 for [v]). Analysis of modulating signals revealed weak modulation at the second and third harmonics. Mean m across all fluent-speech fricatives was $\sim$0.35.

AM in noise at these depths is easily detectable in broadband AM noise carriers. However, research suggests that the combination of a short-duration noise carrier (60–100 ms) and low-frequency tone with f_0 equal to that of the modulation presents a more complex perceptual scenario. In relation to the interaction of AM and tone, it has been suggested that *improvement*, as well as deterioration, in detection is possible.

A series of psychoacoustic experiments was designed to measure thresholds for AM detection in synthetic signals with the same combination of sources as VFs: a wideband-noise carrier, modulated at 125 Hz, and tone of the same frequency. Stimuli were then gradually manipulated, bringing them closer to real VFs, by incorporating more experimental variables. The final experiment used engineered speech stimuli.

AM detection when the tone was less then 10 dB above noise level was predictable. As the tone increased in volume, detection depended on phase: out-of-phase and in-phase stimuli produced optimum detection impedance and enhancement respectively, although the magnitude of the enhancement effect was limited. Little effect for spectral shape was found, but a sharp increase in threshold for shorter stimuli echoed previous ndings. The effect of the phase difference between tone and noise envelope was shown to remain effective for stimuli as short as 60 ms. The final experiment showed that detection of AM is hindered when the the vowel preceding the frication is loud and heard as speech rather than as a tone.

The final stage of the research was to establish whether listeners could use AM as a perceptual cue. In a cue-trading experiment, the phonological voicing boundary was measured along a formant-transition-duration continuum, as a function of AM depth and phase, voicing amplitude and masking of the voicing component. The presence of AM as well as its perceptual prominence were found to trade with formant-transition cues and voicing amplitude. In general, AM increased voiced responses by approximately 30% over the condition with unmodulated frication noise.

Key words: Speech, Perception, Acoustic Phonetics, Fricatives, Amplitude Modulation, Voicing

Email: jon@pincas.co.uk

WWW: http://www.ee.surrey.ac.uk/personal/j.pincas

Acknowledgements

I owe this thesis to Phil Jackson, without who (or should it be whom?) it would certainly never have happened. Twice. Aside from the countless hours of work he has put into the research over the course of the last 5 years, he has helped me see straight many times when my outlook was, well, skewed. Thanks Phil.

Thanks also to...

Francis Rumsey, who has been a willing and able co-supervisor. I appreciated all your words of wisdom.

The URS and whoever was behind it. Money is always useful. While I'm here, I better thank Mum, Pan, Amor and Paco - unfortunately the URS was only 3 years! Of course, your contribution went far beyond financial.

Chris Darwin for being a very hospitable host in Brighton in Summer 2005 and generously affording me access to the University's facilities.

David Or (although it's been so long now, I can't really remember which conjunction your surname was - sorry!) for being the very first subject.

The many other subjects of all the experiments, especially the ones I promised money to but never got round to paying, and the ladies who were forced to produce aerodynamically impossible sounds at 125 Hz. They will tell you that this is very low for the female vocal apparatus. I forgive your octave errors.

Martin, Vina and Jack - I know I wasn't around much, so thanks for defending my desk against invading Spaniards.

Jessi, for putting up with this way longer than any sane person would.

This thesis is dedicated to Bill, Marge, Mary, Andrew, Jaques, Polly and James.

Contents

List of Figures

List of Tables

List of Acronyms

AM	Amplitude modulation
VF	Voiced fricative
SPL	Sound pressure level
VT	Vocal tract
/VF/	Vowel-Fricative
/FV/	Fricative-Vowel
/VFV/	Vowel-Fricative-Vowel
PSHF	Pitch Scaled Harmonic Filter
TMTF	Temporal Modulation Transfer Function
TMN	Tone-masking-noise
NMT	Noise-masking-tone
CMR	Comodulation masking release
SAM	Sinusoidally amplitude-modulated
IRN	Iterative ripple noise
ANOVA	Analysis of Variance (statistical test)
VOT	Voice onset time
SNR	Signal-to-noise ratio
2 or 3AFC	2/3-alternative forced-choice (experimental paradigm)
EGG	Electroglottograph
HP	High-pass (filter)
PAR	Periodic-to-aperiodic ratio
RMS	Root mean square
TNR	Tone-to-noise ratio
CI	Confidence interval

List of Symbols

Speech and Fricatives

f_0	Fundamental frequency
F_{ON}	Frication onset point
F_{OFF}	Frication offset point
F_{rms}	RMS amplitude of frication noise
F1,F2,F3	Numbered formants in speech
h	Harmonic index (in hf_0: $1 = f_0$, $2 =$ first harmonic, etc.)
H	Total number of harmonics
λ	Fricative noise duration
PP	Formant transition duration in pitch periods
T_{ON}	Formant transition start
V_{ON}	Voicing onset point
V_{OFF}	Voicing offset point
v	Voicing strength
$\hat{v}$	*Estimated* voicing strength from measurement procedure
v_h	Voicing strength at hf_0

Signals and Signal Processing

f_s	Sampling frequency
f_{HP}	High-pass filter cutoff frequency
f_{BP}	Band-pass filter cutoff frequencies
f_{BW}	Signal bandwidth
k_h	Fourier transform frequency bin that contains hf_0
$\tilde{k}_h$	Contiguous set of frequency bins under k_h spike and above noise floor
$\bar{X}(k)$	Modulation spectrum
$h_(n)$	Raised cosine ramp
$w(n)$	Noise carrier signal
$x(n)$	Amplitude modulated noise signal
Sc	Scaling factor

Modulation

$a(n)$	Modulating signal
d	Amplitude of modulating sinusoid

f_m	Modulating frequency
m	Modulation depth in standard index form
$\hat{m}$	*Estimated* modulation index from measurement procedure
m_h	Modulation index at hf_0
m_d	AM detection threshold
ϕ	Modulation phase shift or difference
ϕ_h	Modulation phase shift or difference at hf_0

Jets, Fluids and Flows

A	Cross-sectional area of a constriction
D	Diameter of constriction/jet/channel
F_{dB}	Difference between maximum and minimum turbulence sound pressures
μ	Dynamic viscosity of fluid
ΔP_C	Pressure drop across a constriction
p_s	Sound pressure
Re	Reynolds number
ρ	Fluid density
St	Strouhal number
U	Volume velocity flow in cm^3/second
U_e	Mean volume velocity of flow
V	Flow velocity

Stimulus Generation

$C(n)$	Pre-final-stage stimulus resulting from concatenation of engineered components
$F(n)$	Pre-recorded fricative noise as input to stimulus generation procedure
$g(n)$	Signal during gaps between stimulus intervals
$P(n)$	Pre-recorded audio signal as input to stimulus generation procedure
R	Decibel ratio of RMS amplitude of tone to noise, also termed TNR
$s(n)$	Final engineered stimulus interval
$V(n)$	Voicing signal as input to stimulis generation procedure
$y_i(n)$	Synthesised 3-interval trial, subscript corresponds to interval of modulated target

Units

dB	Decibels
Hz	Frequency in Hertz
ms	Milliseconds
Pa	Pascals

Chapter 1

Introduction

This thesis is concerned with elucidating the role of amplitude modulation (AM) in the perception of the voicing distinction in fricatives.

AM is commonly talked about with reference to speech intelligibility. Research in the areas of modulation spectrograms (e.g., Tchorz and Kollmeier (2002)), signal processing techniques for cochlear implants (e.g., Rosen et al. (1999)) and temporal aspects of speech intelligibility (e.g., Shannon et al. (1995)), have demonstrated how 'slow' amplitude modulations (< 50 Hz) provide important information to listeners (especially those with abnormal hearing) and can transmit aspects of the speech signal even when spectral detail is lacking. However, the modulation frequencies of interest in this study are much higher (> 100Hz) and the phenomenon we are seeking to observe is fundamentally different. This work is concerned with the role of 'fast' modulations as a phonetic cue, rather than how slower AM can provide syllable-level detail about an utterance. It is therefore of paramount importance to distinguish the work carried out here from those areas of research, as the two are largely unconnected.

1.1 Background

1.1.1 Fricatives

Human speech is characterised by a rapid alternation between the open vocal tract (VT) configurations of vowels and the closed configurations of consonants. Stop consonants are the canonical example of the latter, where complete occlusion momentarily prevents airflow through the VT, causing a brief silence. Approximant consonants are more vowel-like, involving only a brief gesture in which articulators are brought closer together, but maintained separated enough to allow air to flow freely through the VT.

Fricative sounds in speech bridge these two classes: turbulence noise is produced by forcing air through a constriction in the oral cavity just narrow enough to generate noise within the jet and, more importantly, at or along a physical obstacle downstream of the constriction (Shadle, 1985).

The IPA categorises fricatives by place of articulation and as either voiced or voiceless (International Phonetic Association, 1999). Although this is the standard way of classifying fricatives, other theoretical classification systems that use different feature sets have been proposed (Ladefoged, 2008; Laver, 1994). Although the reality of production complicates matters somewhat[1], the IPA place classification refers to the location of the oral constriction. The binary (voiced/voiceless) phonological voicing distinction corresponds primarily to the presence or absence of vocal chord vibration and accompanying low-frequency energy during frication. English has voiceless and voiced fricatives (VFs) at four places: postalveolar /ʃ,ʒ/, alveolar /s,z/, dental /θ,ð/ and labiodental /f,v/.

1.1.2 Voiced Frication

Voiced frication results when a glottal sound source is present along with the primary turbulence noise source, producing the familiar 'buzzy' quality of mixed excitation. The characteristics of voiced frication do not, however, arise simply from the linear combination of these independent sound sources. The articulatory, aerodynamic and acoustic conditions required by and resulting from the simultaneous production of glottal vibration and frication noise produce 'mutual interaction effects' (Pincas and Jackson, 2004) the presence of each source causes the other to be changed in character from the case where it occurs in isolation. The focus of this thesis, amplitude modulation (AM) of the frication component, is one such effect, in which vocal chord vibration during frication production causes regular, short-term fluctuations in the RMS amplitude of the noise. Other effects include mutual amplitude reduction (Stevens, 1971), changes in fundamental frequency of voicing (Lofqvist et al., 1989), and spectral changes both in the voicing component before, during and after frication (Lofqvist et al., 1995) and in the frication-noise component (Shadle, 1995).

[1]The problems raised by the classification of fricative place of articulation are numerous and their treatment is beyond the scope of this thesis. In summary, whereas theory assigns place of articulation labels solely in terms of the constriction location, fricative sound generation may take place at multiple or spread out locations (such as the palatal [ç]) along the vocal tract, with different generation mechanisms operational at the constriction and/or downstream obstacles. In other cases, the shape of the tongue, rather than its ultimate position, may be a better distinguishing factor between two fricatives (such as [s] versus [ʃ]).

1.1.3 Defining Amplitude Modulation

In AM, a carrier signal $w(n)$ is multiplied by a modulating signal $a(n)$ to produce the amplitude-modulated signal, $x(n) = w(n)a(n)$. With a periodic modulating signal, $a(n)$ takes the form of a d.c. term and fundamental sinusoid of frequency f_0 plus harmonics:

$$x(n) = w(n)\left[1 + \sum_{h=1}^{H} m_h \cos\left(\frac{2\pi h f_0 n}{f_S} + \phi_h\right)\right], \tag{1.1}$$

where $h \in 1..H$ are the fundamental and associated harmonics, m_h is the modulation index at hf_0, f_S is the discrete signal's sampling frequency and ϕ_h is an arbitrary phase shift assumed to be constant. We assume $a(n)$ to be non-negative. With purely sinusoidal amplitude modulation ($H = 1$), $a(n)$ is completely specified by the f_0 component, i.e., by m_1 and ϕ_1. In natural VFs, the noise $w(n)$ is coloured and the underlying modulation shape $a(n)$ is not a pure sinusoid.

Throughout the remainder of this thesis, 'AM' or 'modulation index' without further qualification always refers to modulation depth at f_0 (i.e., m_1) and is designated simply as m. Whilst m is perceptually the most important acoustic variable relating to modulation[2], Equation 1.1 highlights the possibility of modulation at harmonics of f_0 as well as the modulation's phase relationship to the voicing component as parameters of AM.

m can be conceptualised as the fraction of the carrier signal by which the modulated signal varies, e.g., if $m = 0.5$, then the signal fluctuates by 50% above and below its original, unmodulated value. In engineering and signal processing contexts, m is given in standard index form (in the range 0–1). In the perceptual literature, modulation detection/discrimination thresholds are quoted in dB as $20\log_{10}(m)$. Figure 1.1 exemplifies unmodulated broadband noise (8 kHz bandwidth), and noise modulated with three different values of m (f_0 fixed at 150 Hz) within this range (signals based on Equation 1.1). At $m = 0$, there is no systematic fluctuation in the amplitude of the noise. As m increases towards 1, the noise becomes more concentrated into 'puffs', until, at $m = 1$, the noise is 'completely' modulated and almost disappears between bursts.

1.1.4 Detection of Amplitude Modulation

The acuity of AM detection is described by the smallest amount of modulation necessary to successfully distinguish a modulated sound from an identical, unmodulated sound,

[2]Perception of AM becomes progressively more difficult as frequency increases (see 'TMTF' literature, Section 2.2.1).

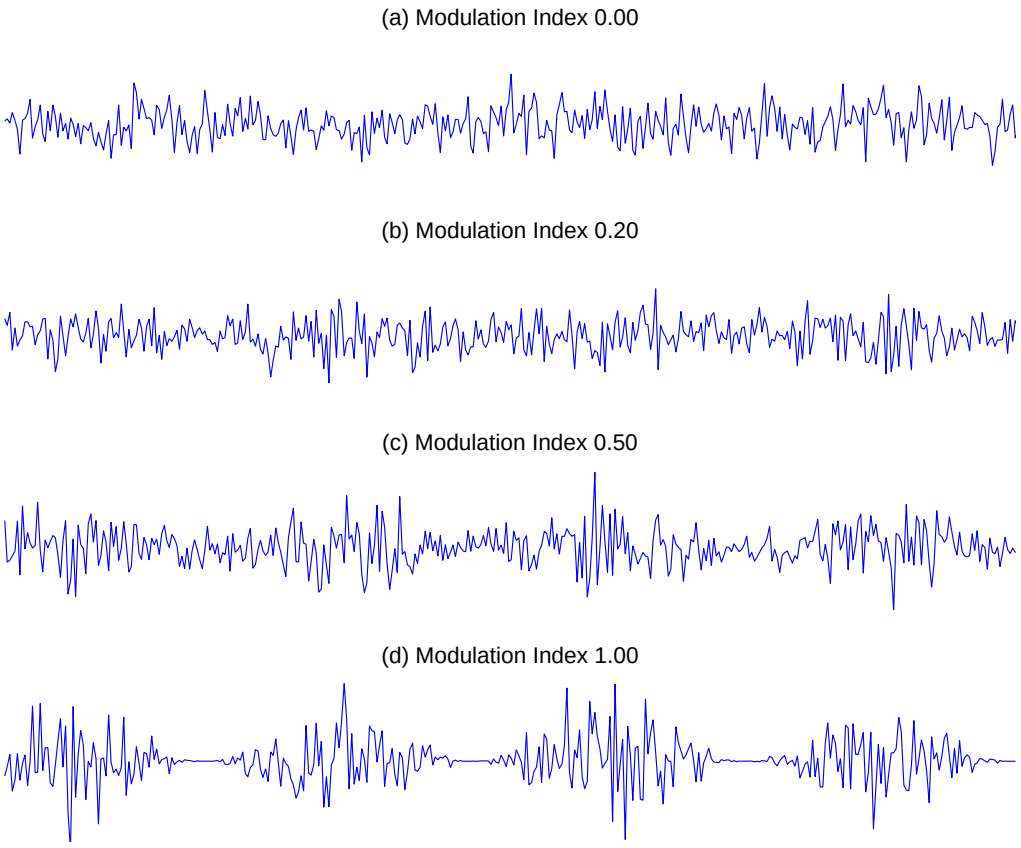

Figure 1.1: Unmodulated and modulated 8 kHz broadband noise. 150 Hz modulation frequency. m values are (a) 0; (b)0.2; (c) 0.5 and (d) 1.

for a given condition. For AM in simple noise carrier stimuli [3], AM-detection thresholds, m_d, have been well studied and for frequencies spanning the possible range of speech fundamentals (i.e., 80 Hz $\leq f_m <$ 400 Hz), are in the range -24 dB ($m_d = 0.06$) to -12 ($m_d = 0.25$), with studies separated by at most 3 dB (Zwicker and Feldtkeller, 1967; Dubrovskii and Tumarkina, 1967; Viemeister, 1973, 1977; Rodenburg, 1972, 1977).

1.2 The Research Problem

Amplitude modulation of the fricative component has long been recognised as an acoustic feature of VFs (Fant, 1960; Stevens, 1998) and has been used the improve the quality of synthesised speech in both voiced frication and aspiration noise (Klatt, 1980; Hermes, 1991) . However, beyond enhancing quality and naturalness, the role of AM as a

[3] Generally 500 ms of broadband noise.

cue to perception of the voicing distinction has received little attention[4]. It is thus not known whether AM plays a role as a cue the voicing distinction. If it does, we should like to know to what extent, and how it interacts with other previously cited cues. If not, we should like to know why.

As suggested in Section 1.1.1, the primary cue to the voicing distinction in fricatives, as with other speech sounds, is considered to be the presence or absence of voicing during frication. In practice, however, the articulatory and aerodynamic vocal tract configuration necessary to produce simultaneous voicing and frication mean that the phonetics do not always reflect phonology. Studies have shown that phonologically voiceless fricatives are often voiced through a substantial part of their duration (Pincas, 2004; Heid and Hawkins, 1999). Conversely, phonologically VFs are often devoiced (Haggard, 1978; Smith, 1997, 1995). In general, simple absence or presence of voicing during frication is not adequate to specify phonological voicing — further measures are necessary.

A number of other possible cues to the voicing distinction have been suggested in the literature: overall segment duration (VFs are shorter than voiceless, e.g., Crystal and House (1988)); length of preceding vowel (Denes, 1955); for intervocalic and word-final fricatives, duration of frication/voicing overlap at segment onset (Pirello et al., 1997), duration of voiceless portion before offset (Pincas, 2004), and duration and gradient of formant transitions at onset and offset (Stevens et al., 1992); and for word-initial fricatives, duration of voiceless portion at onset (Massaro and Cohen, 1976). Most of these variables have been subject to acoustic classification and perceptual experiments revealing varying degrees of reliability as cues. There is still little consensus however, as to which cue, if any, are primary.

AM might be considered a redundant cue in that it is always accompanied by voicing, traditionally considered the primary cue for voicing in fricatives. However, listeners are known to take advantage of multiple sources of information in the classification of speech sounds, including 'redundant cues' (Stevens et al. (1992) is a good example for voicing in fricatives). Theorists have taken perceptual integration of this wide variety of diverse cues as strong evidence in favour of either an articulatory mode of speech perception (Repp and Liberman, 1990), or an extensive, knowledge-based, algorithmic approach in which the acoustic to abstract phonological mapping is made directly (Diehl and Kluender, 1990).

Assigning importance to all possible perceptual cues arising from an articulation leads to a reformulation, or at least reemphasis, of the research question. Rather than seeking

[4]As far as I am aware, only Strope and Alwan (2001) have previously directly considered AM as a possible cue in VFs

invariant, 'primary' cues to perception, we should look to understand how cues are integrated under different circumstances. The consideration of realistic or sub-optimal speech transmission conditions is then brought to the forefront, instead of being treated as an independent field:

> *In designing a practical speech-recognition device, one attempts to maximise accuracy and speech of performance within the limits imposed by the available computational resources. Because these limits are usually quite severe, a reasonable strategy might be to process only a few of the most robust cues for each category and to disregard all the rest. In the case of the human listener, limits on attentional capacity might well dictate this kind of strategy. Why then do listeners seem to use all cues even those of presumably marginal importance? The most likely answer is that a high level of redundancy is required to ensure accuracy of recognition in the general case where aspects of the communications situation are non-optimal. In natural settings, listeners have to contend with noise and reverberation, while in more artificial situations the speech signal may be degraded by a variety of kinds of filtering and distortion. The vocal apparatus, and especially the auditory system, may have defective components that reduce the quality of speech communication. Wide variation in vocal-tract characteristics and dialect make the listeners task even more difficult. A strategy of making full use of redundant cues may be the only way to overcome the impressive odds against successful speech recognition.* Diehl and Kluender (1990)

Sub-optimal conditions where AM of frication noise could aid in the perceptual task are clear: one can envisage a situation where the voicing source is masked by a low-frequency sound (rumble from an external source) or cut off by a low-quality speech communication or reproduction system (such as low-bit-rate coding or poor quality loudspeakers). In these situations, AM might even act as a primary cue to the voicing distinction.

In summary, the research at hand is primarily important to the specific study of voiced fricatives. An understanding of the perceptual role of AM as a cue to the voicing distinction is key to a complete understanding of cue integration and perception of this basic phonological boundary. This is particularly so given the nature of AM as a previously unknown cue. Rather than refining knowledge of a well studied cue, this thesis posits a novel one. Of course, the discovery of a new cue would have important implications in speech technology; recognition and synthesis might take advantage of the AM for quality and accuracy enhancements. In the wider field of speech perception, the study of AM as a 'redundant' cue sheds light on the remarkable ability of listeners

to integrate disparate cues to perception and helps build a more accurate overall picture of which acoustic properties can be successfully integrated.

1.3 Approach and Organisation of the Thesis

1.3.1 Acoustic Study

One of the barriers to the investigation of AM as a perceptual cue is the lack of knowledge regarding its acoustic tendencies or perceptual salience. We have sought to address this issue by conducting a systematic study of AM in fricatives, incorporating acoustic measurements, psychoacoustic AM detection testing and finally, categorical perception-type phonetic testing.

Since little was previously known regarding the distribution of AM in VFs (previously, only Jackson and Shadle (2000) have presented rudimentary data), a first step was to establish its acoustic characteristics and distributions across as many real fricative tokens as possible.

Since modulated frication does not occur in isolation, it was necessary to contemplate various preprocessing techniques before the application of a specifically designed algorithm to extract measures of m for the fundamental and first two harmonics from fricatives in tokens taken from corpora designed and recorded especially for the task. Chapter 3 details these techniques along with results correlating m values across fricatives to a range of secondary variables.

1.3.2 Perceptual Study

There are many ways in which real VFs differ from the controlled stimuli used in AM detection experiments: time-varying spectral shaping of the high-frequency aperiodic component rather than simple broadband or narrowband noise (Shadle, 1995); an overall duration that is much shorter than the 500 ms typically used in threshold experiments (Pincas, 2004); the presence of a low-frequency spectral component (voicing) whose f_0 is equal to that of the modulation frequency, raising the possibility of interaction between spectral and envelope domains; voicing component is complex rather than pure tone, has formants and harmonics, jitter and shimmer with variable f_0; and finally, the fact that they occur in the context of speech rather than in isolation. Some of these acoustic properties can be observed in Figure 1.2, which shows the transition from a vowel into a VF. At transition, the formants move and fade; meanwhile, the high-frequency noise becomes prominent during the fricative segment.

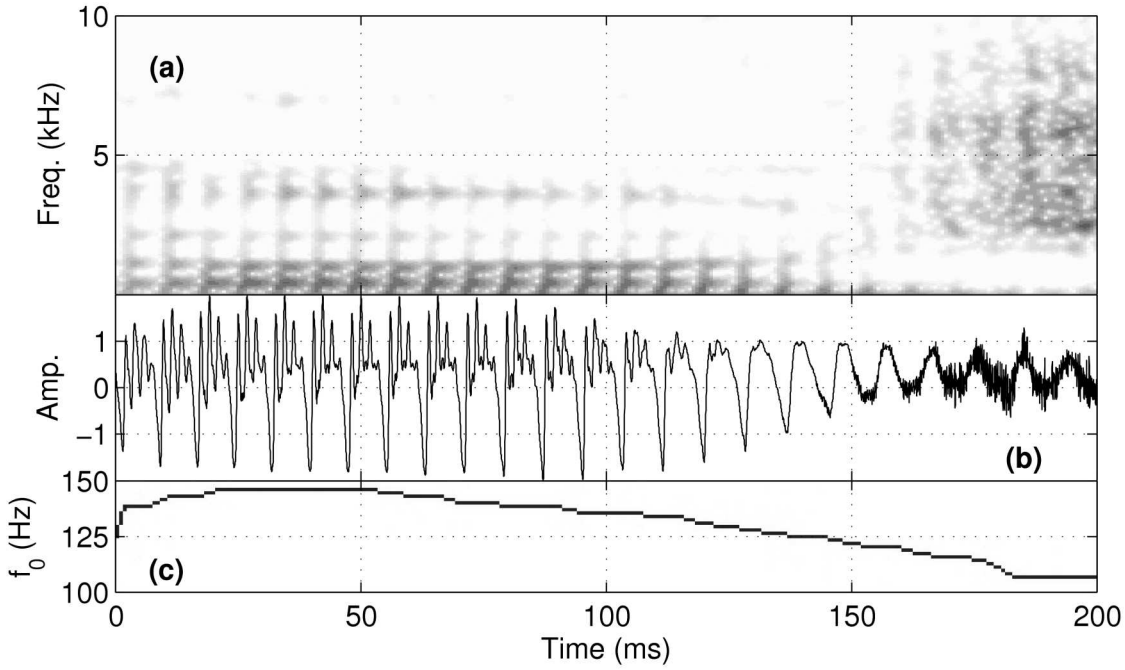

Figure 1.2: (a) Spectrogram, (b) Waveform, and (c) Pitch track of /VF/ transition in [ɑːʒ] token taken from fluent fricative corpus. 16 kHz bandwidth.

Research from different domains within psychoacoustics shows that any of these acoustic characteristics could lead to profound effects on AM detection, e.g., (Stein et al., 2005a; Wakefield and Viemeister, 1985; Wiegrebe and Patterson, 1999). Crucially, this can be impairment or *enhancement*. Whilst this research is now well established, it has not yet been applied specifically to the case of VFs.

The question remains, therefore, as to whether AM in VFs is perceptible and if so, to what extent? If not, which particular acoustic feature of VFs inhibits detection? With a good understanding of AM distribution in fricatives gained from the acoustic study, a thorough psychoacoustic investigation was thus called for.

A series of psychoacoustic AM detection threshold experiments was designed and carried out using stimuli progressing from crude simulations of VFs (simple noise-plus-tone signals) to realistic manipulated-speech stimuli, thus overcoming the limitations of the existing research mentioned above. Furthermore, this progressive approach would isolate the responsible acoustic feature should AM prove to be *undetectable* after any particular manipulation of the stimuli. Finally, the completion of a detailed psychoacoustic study before any speech perception testing would preclude the simple explanation that subjects 'simply cannot hear the AM in the the event that AM was proven to play no role in cueing the voicing distinction, an explanation that would raise more research questions than resolve.

Detailed methodology and results for these experiments appear in Chapter 4.

1.3.3 Phonetic Study

Finally, a phonetic cue-trading experiment was designed to investigate whether subjects were able to use AM as a cue to the voicing distinction. Casual listening and pilot tests revealed that a formant transition continuum was optimum at eliciting 100% correct voicing responses at the extremes with a sharp transition between voicing state responses around the centre point of the continuum [5]. Amplitude modulation was then introduced, with variations in depth and phase relationship to the voicing. In a further variation, a low-frequency rumble was introduced to mask voicing. Chapter 5 details the method used and results of this experiment.

[5] A sharp transition between responses at a particular point along the continuum is indicative of a phonetic boundary effect (Pickett, 1999)

Chapter 2

Literature Review

2.1 Acoustic Study

There is little existing data from acoustic study of AM in VFs; in fact, only Jackson and Shadle (2000) have previously made direct measurements to quantify AM depth or phase. This results of this study are covered in Section 2.1.1.

Also considered are results from statistical and modelling work in aerodynamics and speech production. In some cases, studies yield quantitative predictions against which the results of this work can be compared. In other cases, although they may not directly predict m, modelling studies provide insight into the range of physical aerodynamic and gestural parameters that dictate AM characteristics in real fricative tokens.

In either case, predictions of m and the parameters that control it can only be tentative estimates given the complexity and apparent lack of agreement regarding the mechanism(s) responsible for the generation of AM-noise in the vocal tract. These mechanisms and their predictions with respect to the acoustics of AM are covered in Section 2.1.2.

2.1.1 Acoustic Measurement of Speech

Jackson and Shadle (2000) published limited data relating to amplitude and phase of modulation in various fricatives. Ten fricative tokens were used with seven places of articulation. A signal processing tool – the Pitch Scaled Harmonic Filter (PSHF) — was used to separate harmonic and anharmonic streams (of which the latter is of interest here). Oscillations in short-term power of the anharmonic component were quantified by measuring the strength of the harmonic relating to the fundamental. Figure 2.1 illustrates their results for magnitude of modulation across different voiceless fricatives. As can be seen, modulation was found to range from 0 dB in the case of [ß] to 2 dB ($m \approx 0.25$) in the case of [z]; modulation for the other fricatives tends to cluster around 1 dB ($m \approx 0.1$).

Modulation phase (ϕ) was reported with reference to an Electroglottograph signal. Results indicate an approximate 90° phase separation between bilabial ([β]) and pharyngeal ([ʕ]) VFs with a gradual phase transition as the point of constriction moves in a posterior direction from former to latter.

2.1.2 AM Generation Mechanisms

During voiced frication, transglottal pressure and laryngeal tension maintain phonation. Glottal vibration causes AM indirectly though variations in flow through the

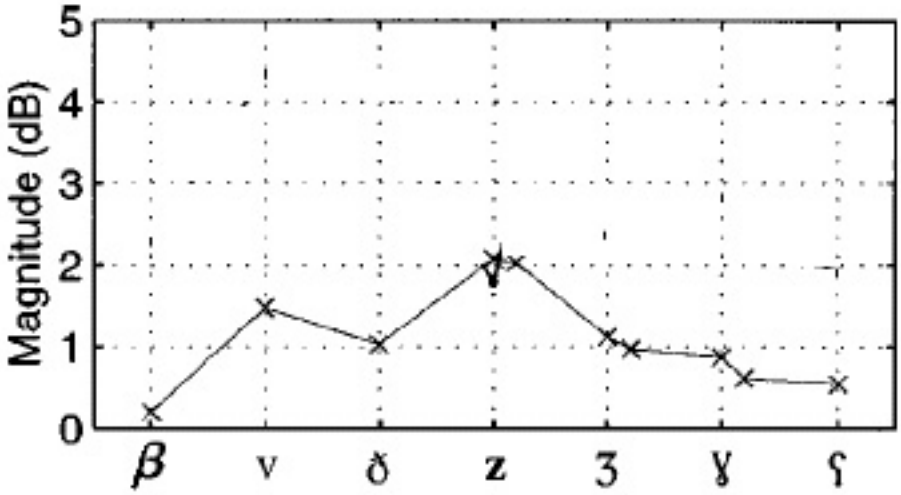

Figure 2.1: Magnitude of modulation coefficients versus place of articulation for sustained fricatives. Measurements on vertical grid lines are for normal voicing; those adjacent where a pair of measurements are shown were taken from a section interrupted by devoicing. Source: Jackson and Shadle (2000).

constriction, assumed to be fixed (Shadle, 1985). Although the presence of AM noise in VFs is broadly acknowledged, fundamental questions remain. How does the result of glottal vibration reach the constriction? How does this perturbation affect the jet and the turbulence it forms? How does the perturbed turbulence go on to generate sound?

The Static Approach: Description

In his view of speech production, Stevens (1971) models noise sources under *static* aerodynamic conditions. From this perspective, the opening and closing of the vocal

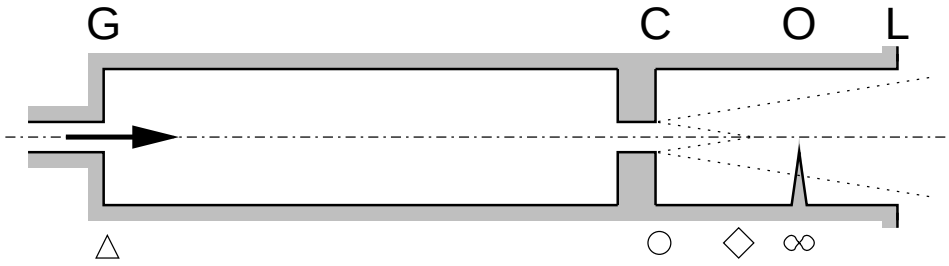

Figure 2.2: Sound production mechanisms in a schematic mid-sagittal view of the vocal tract in VF configuration: (G)lottis, (C)onstriction, (O)bstacle, (L)ip termination. Acoustic sources: △ periodic, ∘ monopole, ♢ quadrupole, and ∞ dipole noise.

folds during one glottal cycle turns the flow on and off.

Acoustic theory of noise generation from turbulence describes three types of source: monopole, dipole and quadrupole (Lighthill, 1952), illustrated in Figure 2.2 . Monopoles arise from velocity fluctuations injected into the sound field, dipole sources result from turbulent flow impinging on a solid obstacle, and quadrupoles occur in regions of turbulence through self mixing. The intensity of these sources depends on the flow velocity as V^4, V^6 and V^8 respectively (Lighthill, 1954). In fricatives, flow at the constriction exit produces a monopole, then quadrupoles just downstream in the jet core and dipoles at the teeth or lips (Stevens, 1998). It is widely accepted that dipole sources dominate noise generation in fricatives (Stevens, 1971; Shadle, 1985, 1990), although some studies have considered a monopole component (Pastel, 1987; Stevens, 1998; Narayanan and Alwan, 2000).

Thus, the power of dipole sources is modulated proportional to V^6 while the flow velocity depends on the area and pressure across the constriction, ΔP_C (other sources accordingly). All these changes are considered to happen simultaneously.

Predictions from this model depend largely on sub and supraglottal constriction sizes, giving different modulation possibilities for different fricatives. When the area of the supraglottal constriction is large, the combined pressure drop in the vocal tract is mainly dictated by the area of the glottal constriction and thus modulation of the airflow is deeper than when the supraglottal constriction is small (and therefore responsible for the majority of the pressure drop). The nature of the transition from the constriction to the post-constriction cavity will also affect the degree of pressure drop across the constriction. Abrupt transitions, such as those they found for [s], induce greater losses and thus a bigger pressure drop than for smoother transitions, as in the case of [ʃ] (Narayanan, 1995; Narayanan et al., 1995).

A few studies have reported constriction dimensions for fricatives based on direct measurement from MRI scans (Narayanan et al., 1995) or from indirectly calculating the area from aerodynamic measurements using the 'orifice equation' (Beautemps et al., 1995; Scully et al., 1992; Badin and Fant, 1989) [1]. These studies report constriction areas in the approximate range $0.05-0.3$ cm^2, with an impressionistic average around 0.15 cm^2, although no systematic trend for place that is consistent across the studies can be identified. Stevens (1971) concludes, based on calculations using airflow and pressure measurements from four other studies[2], that supraglottal constriction lies in the range $0.05-0.2$ cm^2.

Assuming the latter as a conservative estimate of the range of constriction areas, a

[1] Badin and Fant (1989) used a different one to Beautemps et al. (1995).

[2] Malécot (1955); Isshiki and Ringel (1964); Hixon (1966); Arkebauer et al. (1967).

glottal area varying between approximately zero and 0.3 cm^2 during phonation and a subglottal pressure of 8 cm H_20, it is predicted that the smallest fricative constriction produces 25 dB less airflow modulation[3] than the largest (Stevens, 1971). Based on single dipole sound pressure (p_s) model of noise generation, governed by

$$p_s = KU^3A^{-2.5}, \tag{2.1}$$

where U is volume velocity flow (cm^3/sec), A is the cross-sectional area of the constriction (which, all other factors remaining the same, affects the area of the surface over which turbulence occurs), and K is a constant influenced by the exact configuration of downstream obstacles, this formulation predicts an approximate difference between maximum and minimum turbulence sound pressure at the supraglottal constriction during phonation (F_{dB}) of 10 dB for the smallest constriction considered above (0.05 cm^2) and 20 dB for the largest (0.2 cm^2). Modulation depth, m, can then be calculated from F_{dB} as follows:

$$\begin{aligned} F_{dB} &= 20\log_{10}\left(\frac{P_{max}}{P_{min}}\right) \\ &= 20\log_{10}\left(\frac{1+m}{1+m}\right) \\ m &= \frac{10^{\frac{F_{dB}}{20}}-1}{1+10^{\frac{F_{dB}}{20}}}, \end{aligned} \tag{2.2}$$

where P_{max} and P_{min} are the maximum and minimum sound pressures. Hence in terms of modulation index, the predicted range is $0.5 -- 0.8$ depending on constriction size. Stevens (1971) exact prediction of maximum modulation, based on slightly different estimates of constriction sizes is 15 dB ($m \approx 0.7$).

Beyond the The Static Approach

Possible alternative views of AM noise generation highlight periodicity in the flow either before or after the fricative constriction. Turbulent flows often exhibit large-scale regularity at certain ranges of flow rate and Reynolds number, $\mathrm{Re} = \rho VD/\mu$, (Sinder, 1999). In fact, periodicity in unstable fluid flows is a well attested phenomenon in fluid mechanics (Massey and Smith, 1998; Tritton, 1988; Munson et al., 1990): for example, a periodic flow develops downstream of a cylinder placed in a moving fluid due to the formation of a regular pattern of vortices being shed. This is known as a Kármán

[3]Difference between peak and minimum airflows duration a single cycle of phonation.

vortex trail (Munson et al., 1990); it is the phenomenon responsible for the 'singing' of electricity or telephone cables in the wind and the instability of bridges when shedding frequency coincides with the natural frequency of the structure. The frequency of vortex shedding, f_0, is usually characterised in terms of the dimensionless Strouhal number, St, which is given by $f_0 D/V$, where D is the exit diameter of the constriction and V is the flow velocity. For moderate Reynolds numbers ($250 < \mathrm{Re} < 2 \times 10^5$), a Strouhal number of 0.2 is typical (Tritton, 1988; Massey and Smith, 1998).

Crow and Champagne (1971) studied the development of jets issuing from a cylindrical constriction: flow visualisation experiments with water showed how instability in jet formation can transform gradually into regularity at increasing Reynolds numbers, as illustrated in Figure 2.3. The first sign of instability is a slight 'whiplash' motion in the stream (a). As the Reynolds number increases, the jet shape progresses to a 'corkscrew', then to 'lobes', and finally, at a Reynolds number of approximately 10^3, to a train of axisymmetric 'puffs'. The Strouhal number characterising the 'train of puffs' was found, by counting, to be approximately 0.3 and largely independent of the Reynolds number.

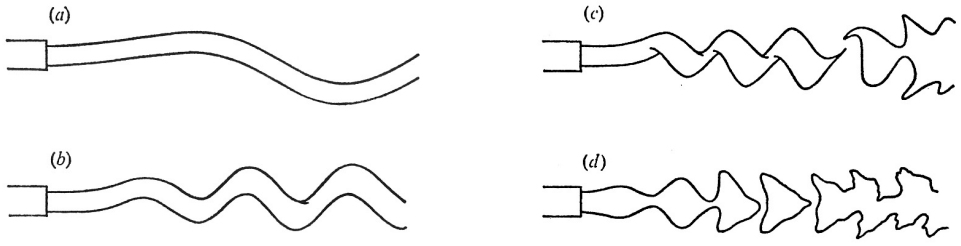

Figure 2.3: Evolution of jet instability with advancing Reynolds number. Jets (a) to (d) span the Reynolds number interval from approximately 10^2 to 10^3. Source: Crow and Champagne (1971).

Furthermore, unstable vortex formation is sensitive to acoustic interference and a sound wave near the jet's natural Strouhal number ($\mathrm{St} = f_0 D/V$) regularises or *forces* the turbulence, causing rotational flow structures to grow periodically (Simcox and Hoglund, 1971; Crow and Champagne, 1971).

Modifying the jet facility used in the original experiment, they included a loudspeaker used to produce a forcing sine wave. Based on the previous finding that the 'preferred' Strouhal number was 0.3, they set about testing the effect of a forcing acoustic signal with a frequency of 185 Hz, with the Reynolds number of the jet set at 10.6×10^4 (yielding $\mathrm{St} = 0.3$). Figure 2.4 shows their results. The x-axis gives the RMS fluctuation amplitude of the forcing signal, u_e (as a fraction of the mean flow U_e); the y-axis gives

the RMS fluctuation of the axial component of jet velocity u in response to this forcing (again, as a fraction of mean flow U_e). The figure shows that response rises almost linearly with small forcing amplitudes and then saturates with higher amplitudes. The biggest changes take place with a forcing amplitude of $0-2\%$ which causes cyclical fluctuation to increase from background level (approximately 4%) to around 17%. No amount of forcing will cause a response of over approximately 19% ($m \sim 0.19$).

Further results reveal a dependency of response on the distance of measurement from the constriction, with a distance equal to four constriction diameters giving optimal response. Finally, there is a complex relationship between response shape and the Strouhal number relating to the forcing frequency: at $\mathrm{St} = 0.3$, the preferred mode of oscillation, response is maximal.

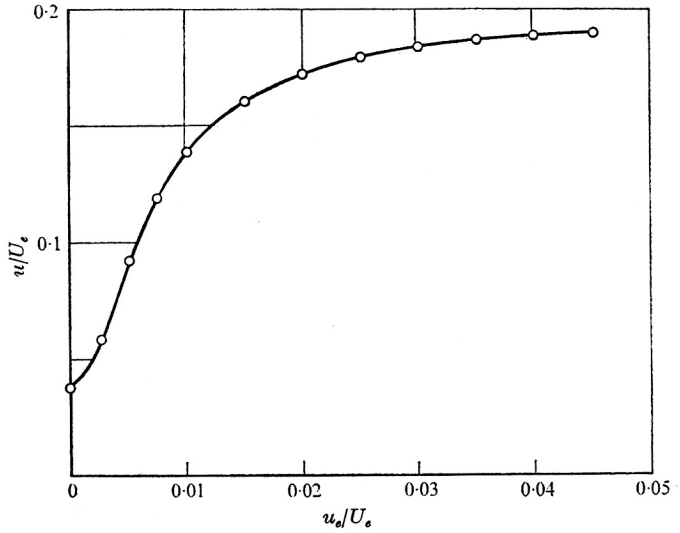

Figure 2.4: Amplitude response at the preferred Strouhal number 0.30, measured 4 jet diameters downstream of the jet exit. u/U_e is the RMS fluctuation amplitude of the noise. u_e/U_e is the RMS fluctuation amplitude of the forcing signal. Source: Crow and Champagne (1971)

In VFs ($0.05 \leq \mathrm{St} \leq 0.2$), it is assumed that voicing sets up the forcing wave which then interacts with the turbulent jet at the fricative constriction to produce periodic vortices that convect downstream. According to this view of modulation generation, these structures modulate the noise generation as the vortex train passes the obstacle (Jackson and Shadle, 2000). In this case, we can predict that the amplitude of the glottal source (amount of forcing) and the geometry of both the constriction and the VT superior to the constriction will play major roles in determining AM depth. Furthermore, the results of Crow and Champagne (1971) would suggest a saturation point

above which no amount of 'forcing' will cause deeper modulation.

A further possible AM generation mechanism comes from glottal vortices causing periodicity in the flow reaching the fricative constriction. Experiments on physical models of the vocal tract with mechanical shutters replicating the action of the vocal folds reveal considerable irregularity near the glottis which changes drastically in character moving downstream (Barney et al., 1999). For turbulent jets, Davies (1981) asserts that the shear layer will form into a vortex train by 20 jet-widths downstream. Measurements made by Barney et al. (1999) suggest that this phenomenon shows up in the vocal tract as a non-acoustic periodicity in the flow at a point distant to the glottis (14 cm in their case) as regular vortices pass[4]. Subtracting their estimate of acoustic particle velocity from the measurement of total velocity fluctuation, velocity appears to vary from approximately -8 cm/s to 6 cm/s.

AM could then occur as these periodicities are translated into frication as self-mixing (quadrupole) noise in the jet or as 'pulses' in the jet impinging on the downstream obstacle. Given the indication that quadrupole (self-mixing) noise sources are less important to fricative noise generation, the former mechanism is assumed to be of little importance. Assuming the latter mechanism and the dipole noise production formulation, a velocity fluctuation of the above quoted magnitude would lead to an estimate of AM of 0.7. Of course, there is no reason to assume that this mechanism functions in isolation from the 'forcing' mechanism previously discussed.

2.1.3 Summary

Systematic measurement of AM in VFs has previously been limited to one study, the results of which suggest a rather low level of modulation ($0 < m < 0.25$) across fricatives.

Mechanisms responsible for AM noise generation in the vocal tract are not well understood, leading to difficulty in modelling the processes and calculating possible AM-depth ranges.

Nonetheless, various approaches and corresponding calculations expounded by Stevens suggest AM in the range ($0.5 < \mathrm{m} ¡ 0.8$) with pressure drop over the supraglottal constriction (dictated by size and shape) as well as the precise noise generation mechanism downstream of the constriction the main determining factors. Furthermore, the strength of voicing, acting as a forcing wave, may determine AM depth, but only to the 'saturation' point, where AM can be forced no further.

[4]This is slightly different from the regular vortex train produced by a single, constant turbulent jet. As the 'jet' formed by glottal opening is only momentarily present, there will be only one vortex per glottal cycle, referred to by Zhao et al. (2000) as the leading vortex, resulting in $11-17$ vortices in the duct at any one time (Barney et al., 1999)

2.2 Psychoacoustic Study

Amplitude-modulated (AM) sounds are a common part of the sonic landscape. Birds, insects and musicians use vibrato to great effect and automotive engineers can diagnose faulty engines by paying close attention to unusual variations in noise. In speech, voicing is often accompanied by aspiration or frication noise with a periodically varying envelope. AM can range from a subtle, inaudible fluctuation of amplitude over time, to a large, momentary variation where instantaneous amplitude is many times greater than the long-term RMS of the sound. Furthermore, AM can be slow, such as the syllable level amplitude changes in speech that occur below 60 Hz, or extremely fast. The auditory system is adept at distinguishing modulated signals from their unmodulated counterparts, even at high frequencies and low modulation depths (i.e., small values of m). Detection of modulation in a signal may be based on explicit sensitivity to amplitude variation, or the effect may be heard as a change in timbre from unmodulated to modulated which is most often qualified as 'roughness' (Zwicker and Fastl, 1999).

In establishing the characteristics of AM detection in voiced fricatives, difficulties arise from the lack of previous study and the complexity of the signal. Although studies have not examined detection in fricatives directly, there is a large body of literature dedicated to AM detection for a multitude of signal permutations. The aim here is to identify which are of relevance to the present research paradigm.

The factors affecting AM detection are thus discussed with particular emphasis on stimuli characteristics relevant to voiced fricatives. AM detection in simple noise stimuli is covered in Section 2.2.1, with reference to stimulus duration and noise properties. In Section , the likely effects of a simultaneous tone (represented by voicing in fricatives) are introduced. Finally, the role of the wider stimulus context (for example, a vowel environment for voiced fricatives) is analysed in Section 2.2.3.

2.2.1 AM Detection in Basic Noise Stimuli

The threshold of AM *detection* is described by the smallest amount of modulation necessary to successfully distinguish a modulated sound from an identical, unmodulated sound, for a given condition (such as modulation frequency, or noise type). Likewise, the threshold of *discrimination* is the smallest difference in AM required to distinguish one modulated sound from another.

The relationship between AM-detection threshold, m_d (most often expressed in dB format as $20\log m_d$), and modulation frequency, f_m, is described by the 'temporal modulation transfer function' (TMTF). The shape of the TMTF has been reported

extensively for broadband noise as well as a range of other carriers, and has been investigated for many permutations of other stimulus variables. Early work on the properties of the TMTF for broadband noise carriers described its general shape as having a low-pass filter characteristic, with precise cut-off frequency and attenuation rate varying slightly across studies (Zwicker and Feldtkeller, 1967; Dubrovskii and Tumarkina, 1967; Viemeister, 1973, 1977; Rodenburg, 1972, 1977). Further work has shown that for $f_m \approx 50-1000$, threshold increases at about 3–4 dB per octave (Viemeister, 1979; Bacon and Viemeister, 1985; Formby and Muir, 1988). Above and below this range, m_d is independent of f_m. Figure 2.5 compares TMTFs obtained by these three studies, focusing on frequencies spanning the possible range of speech fundamentals (i.e., $f_m \approx 50$–400 Hz). Detection thresholds are in the range -24 dB ($m_d = 0.06$) to -12 dB ($m_d = 0.25$), with studies separated by at most 3 dB.

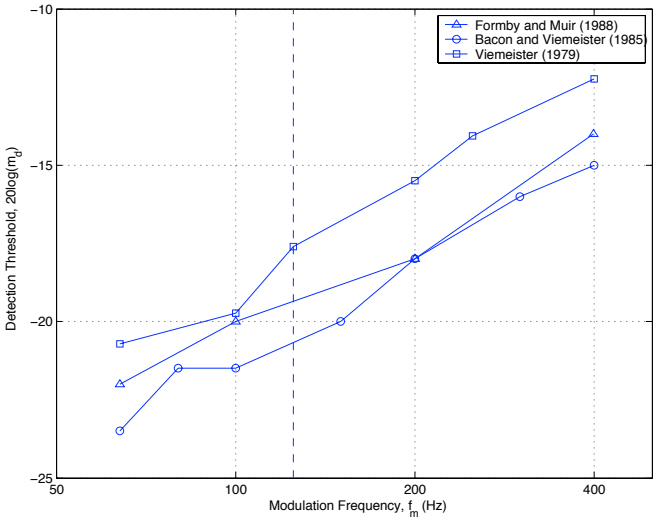

Figure 2.5: Temporal Modulation Transfer Functions for broadband carriers over the frequency range 50 – 400 Hz as reported in 3 previous studies. Triangles: Formby and Muir (1988), Circles: Bacon and Viemeister (1985), and Squares: Viemeister (1979). Dashed line indicates modulation frequency used in this study ($f_m = 125$ Hz).

The slope and intercept of the TMTF for noise carriers has been shown to depend on a multitude of other factors, among them: stimulus duration (Viemeister, 1979; Lee and Bacon, 1997), spectral properties of the noise, e.g., high-pass or low-pass filtering (Formby and Muir, 1988), bandwidth and center frequency (Viemeister, 1979; Strickland and Viemeister, 1997; Dau et al., 1997), overall presentation level (Viemeister, 1979) and the detailed onset pattern of the noise (Viemeister, 1970; Sheft and Yost, 1990). Although the TMTF is affected by many factors, the overall pattern of attenuation at increasing modulation frequencies remains constant.

The Effect of Duration

In both AM detection and discrimination, the difficulty of the task is increased when stimulus duration is decreased. Early work on beats showed detection improvements when increasing stimulus duration from 125 ms to 500 ms (Viemeister, 1970). In a study of AM detection, Viemeister (1979) found that at modulation rates of 64 – 1000 Hz, increase in threshold when decreasing stimulus duration to 250 ms from 1500 ms is only ∼2 – 3 dB. For shorter durations, the results of Sheft and Yost (1990) suggest that the effect of stimulus duration reduction may be more dramatic: between 400 and 6 ms, threshold increased by more than 10 dB. For discrimination tasks, a similar effect is observed for stimulus durations reduced from 800 ms to a duration equivalent to approximately 4 cycles of modulation, after which deterioration in discrimination performance is more marked than for detection (Lee and Bacon, 1997).

The Effect of Spectral and Intensity Properties of the Noise

Formby and Muir (1988) found higher detection thresholds for low-pass filter noise than for broadband or high-pass signals; furthermore, the presence of high-pass content appeared to more important than the signal bandwidth

Viemeister (1979), using narrowband noise carriers centered at 200 Hz, 1 kHz and 10 kHz also found higher thresholds overall for the lower frequency noisebands but attributed this to narrower bandwidths (-3 dB bandwidths were 120 Hz, 500 Hz, and 4.4 kHz respectively). However, an increase in the TMTF cutoff frequency as carrier center frequency increased is attributed to center frequency rather than bandwidth.

Strickland and Viemeister (1997) found no effect on sensitivy or cutoff frequency for specrtal region of the narrowband carrier, but their results appear to confirm that wider bandwidths do improve detection sensitivity.

For very narrow bands of noise, Dau et al. (1997) show a complex relationship between AM detection threshold and modulation frequency as bandwidth increases from 3 Hz to 314 Hz.

Viemeister (1979) also investigated the effect on AM detection of carrier presentation level. For the three highest levels (50, 40 and 20 dB), TMTFs were all within 1 dB of the average. When the level was reduced to 0 dB, AM detection thresholds increased by approximately 3 dB.

AM detection also appears to depend on listener 'adaptation' to modulation which may depend on the manner in which modulation is introduced with relation to the carrier. The shape of the TMTF has been shown to depend on whether modulation is

gated, preceeded by a section of unmodulated carrier, or presented withing a continuous carrier (Viemeister, 1970; Sheft and Yost, 1990).

2.2.2 Perceptual Interaction between Envelope and Spectral Domains

In this study, broadband noise carriers are accompanied by a sinusoidal spectral component at the modulation frequency ($f_0 = f_m$), a condition often observed in real-world sounds such as speech (where f_0 represents the voicing fundamental produced by glottal vibration) and sounds produced by automotive engines (f_0 and harmonics correspond to engine speed and number of cylinders) and industrial machinery. This condition has received comparatively little attention thus far, not withstanding its frequent natural occurrence. In this section, evidence from a broad spectrum of both direct and indirect studies suggesting any possibility perceptual interaction between envelope and spectral domains is discussed and its implications for the detection of AM when accompanied by a tone summarised.

Tone-on-Noise Masking

One way that the presence of a tone might hinder AM detection is through basic auditory masking, in this case the 'Tone-Masking-Noise' paradigm.

Tone-masking-noise (TMN) is a much less commonly studied phenomenon than noise-masking-tone (NMT; see Sporer and Schroder (1992)); this may be partly due the usefulness of the latter paradigm in demonstrating basic psychoacoustic principles, such as the 'Critical Bandwidth' (Fletcher, 1940). Furthermore, TMN is much less effective than NMT; for example, for a 410 Hz tone centred in a 1 Bark noise band, the tone is masked when the spectral level of the noise is just 4 dB above the tone level. For the tone to mask the noise under identical circumstances, the tone level must exceed that of the noise by 24 dB (Pohlmann, 2005).

Thus, the magnitude of the effect would be determined primarily by the loudness of the tone in relation to the noise — as the tone grows in amplitude, the extend of 'upward-spread' masking will reduce the audibility of the noise in the low-frequency region and the restrict availability of the carrier for the AM detection task to the high-frequency region. Note that this high-pass filtering may reduce the modulation depth, although the extent of the effect appears to be small (Eddins, 1993). Furthermore, Viemeister (1979) has shown that AM detection thresholds can rise at low noise-presentation levels which may be relevant in the case of a very loud tone. Note though, that the phenomenon of 'Comodulation Masking Release', or CMR (Verhey et al., 2003), sug-

gests that the presence of modulation in the noise decreases its masking potential (see below).

Perecptial Grouping and Comodulation Masking Release

Amplitude modulation has been suggested as one of the primary 'simultaneous' (as opposed to 'sequential', see Bregman (1990)) cues used by listeners in the grouping of spectrally-disparate auditory features into 'auditory objects': spectral regions or components with similar envelope patterns (or comodulation, based on AM rate and depth) tend to be perceived as coming from the same 'auditory object' (Grimault et al., Mar). This result is noted, for example, in increased intelligibility of comodulated sine-wave speech (Lewis and Carrell, 2007), increased detection of speech in a comodulated background masker (Grose and Hall, 1992), and in'comodulation masking release' (CMR) (Hall et al., 1984).

A typical CMR experiment shows that the masking of a tone by noise is 'released' when the noise is amplitude modulated; furthermore, in contrast to the traditional masking paradigm, masking decreases as bandwidth of the noise increases (Hall et al., 1984). It is unclear how a release from Noise-on-Tone masking might affect the ability of listeners to detect AM (although the possibility is worth considering). Moreover, it has been shown that the CMR effect is reduced at high modulation frquencies, such as those represented by speech fundamentals (Bacon et al., Mar; Bacon and Lee, Jun).

Non-Spectral Pitch

It can be shown that a sensation of pitch, normally associated with periodic frequency-domain components, can be produced by various manipulations of aperiodic stimuli. Subjects can match the fundamental of complex or pure-tone waveforms to the modulation rate of 'interrupted' (Pollack, Jan; Harris, 1963; Miller and Taylor, 1948) or sinusoidally amplitude-modulated (SAM) (Burns and Viemeister, 1976, 1981) noise. This effect is referred to as 'non-spectral pitch' by Burns and Veimeister and can be attributed to the envelope of the interrupted or SAM noise, making it is perceptually akin to a periodic frequency component, at least at some level of auditory processing. Such findings have led to the conclusion that it is more parsimonious to deal with pitch and modulation processing within the same model or framework, e.g., Dau et al. (1997).

The above is of importance to the current study as it suggests that, just as two or more periodic stimuli can interact perceptually (e.g. in masking, leading to enhancement or impairment of detection), the same possibility can be extended to the combination of

periodic stimulus and a 'simulated' or 'non-spectral' tone, such as is represented by the underlying modulation signal in SAM noise.

Temporal Interactions Between Pure Tones and Modulated Noise

Wakefield and Viemeister (1985) looked at detection of AM in noise accompanied by a tone ($f_0 = f_m$) using a 3 kHz noise band centred at 10 kHz. They showed that the effect of the sinusoid was dependent on its level and phase relationship to the noise modulator, with minimum and maximum thresholds separated by 180° - the tone effectively 'masked' the AM signal under certain circumstances, but also, importantly, was able to enhance AM detection when in-phase.

The explanation for this phenomenon put forth by Wakefield and Viemeister (1985) is in terms of additive or multiplicative masking (or enhancement) in the spectral region of the AM carrier (i.e., interaction of the low-frequency tone with the output of auditory filters around 10 kHz). For an additive effect, of the type formalised in Zwicker's 1976 model, they state that "temporal interaction may reflect changes in the envelope at 10 kHz brought about by the addition of the residual energy from the low-frequency tone to the signal at 10 kHz prior to the envelope extraction by the auditory system". A similar explanation appears to have been put forward by McFadden (Apr) to explain how the periodicity in the envelope of a two-tone complex can interact with the 'cycle-by-cycle' periodicity in a low-frequency tone.

For a multiplicative effect, the explanation is that "the low-frequency tone affects the envelope of the high-frequency carrier through modulation of the sensitivity of the 10 kHz channel rather than through addition of energy to the output of that channel". It is argued that either explanation is compatible with their results and neither alters the more important overall observation that "a temporal interaction over such a wide frequency range underscores the fact that auditory interactions can occur in cases for which masking is not observed and may not be expected".

Masking Beteen Fine-Structure and Envelope

In the above study, Wakefield and Viemeister (1985) demonstrated interaction between envelope periodicity and a separate periodic spectral component. Stein et al. (2005b) showed that the same result could also be obtained when the periodicity was present in the carrier, rather than as a separate spectral component. Carrier periodicity was achieved using an 'iterated ripple noise' (IRN) carrier for AM, a broadband noise with significant regular peaks and troughs in the spectrum, producing a strong sensation

of pitch[5]. Throughout the frequency range 8 – 1000 Hz, the presence of this carrier periodicity (again, $f_0 = f_m$) impaired AM detection by $\sim$5 dB compared to a simple broadband carrier.

In a related experiment, McFadden (1988) investigated detectability of a 200 Hz tone in the presence of a tonal complex that gave rise to a periodicity of 200 Hz. The objective was to see if the missing-fundamental waveform contributed to the masking of the tonal signal; however, no such 'periodicity masking' effect was observed.

Distortion Tone

Listeners may take advantage of percepts other than purely temporal cues in the detection of AM in narrowband noise.

Strickland and Viemeister (1997) suggested that bandpass filtered SAM noise produces a distortion tone on the basilar membrane (perceived spectral component) at the modulation frequency.

Wiegrebe and Patterson (1999) confirmed the presence of this tone in casual listening tests and suggested that a sinusoid close in frequency to the distortion tone could produce beats. They then quantified the level and phase of the distortion tone with psychoacoustic AM detection experiments using a 'cancellation' tone. The most important aspect of their work to the current research relates to the relationship between noisebandwidth and AM detection. Their tests were based on bands of noise whose width were twice the modulation frequency (e.g., a 500 Hz bandwidth carrier centred at 4.25 kHz with a modulation frequency of 250 Hz). Setting modulation to 100% and subsequently filtering the carrier to the required bandwidth, they established that AM detection was reduced to chance when the distortion tone was masked.

They conclude that filtering reduces the effective AM depth such that detection based on the temporal cue is no longer possible, but an audible distortion tone is still produced. Importantly, this ratio of carrier bandwidth to modulation frequency at which the temporal cue is not available appears to be a constant rule-of-thumb. They confirm that when the carrier bandwidth is increased to four times the modulation frequency, listeners are consistently able to use the temporal cue (reported as a rattle) to detect AM. Thus, for voiced fricatives which exhibit noise throughout the spectral range, there should be no audible distortion tone to which the periodic voicing component could act as cancellation tone.

[5]The resulting stimulus is termed SAM IRN

2.2.3 The Role of Stimulus Environment

In the discussion of AM detection, the focus so far has been on characteristics inherent to the amplitude modulated noise itself (modulation rate, spectral properties, duration) and on the properties of the stimultaneous tone and its relation to the modulating signal.

In voiced fricatives, AM occurs in a speech context and modulated noise is preceded and/or followed by vowel or consonant speech sounds depending on whether the VF is in word-initial, final or intervocalic position. The spectral and temporal properties of these acoustic environments as well as their amplitude are subject to a large degree of variability depending on the specific word or utterance.

The implications of this variable speech environment for AM detection can potentially be related to a number of psychoacoustic phenomena.

Nonsimultaneous Masking

Nonsimultaneous masking refers to that which occurs when the signal is presented slightly before or after the masker, rather than simultaneous with it. Forward masking, where the signal to be detected follows the masker, is the dominant form, whereas backward masking appears to be dependant the subject's experience (Oxenham and Moore, 1994). The magnitude of a forward masking effect has been shown to depend primarily on the delay between masker offset and probe, with an upper limit of 200 ms beyond which no effect is observed (Moore and Glasberg, 1983). There is also an effect for masker duration, although there is disagreement over the duration beyond which further lengthening of the masker has no effect, with estimates of between 50 ms and 200 ms (Zwicker, 1984; Fastl, 1976). As in simultaneous masking, the spectral properties of both masker and probe as well as their relative intensities also influence the degree of forward masking.

In the case of intervocalic and word-final VFs, forward masking of the modulated noise would be by the immediately preceding vowel and thus the paradigm is the same as for the simultaneous masking discussed in Section 2.2.2 : tone-on-noise. Note that as there is no delay between masker (vowel) and modulated noise, the chief variable of interest is the relative amplitude of the vowel environment.

The Speech Context

At this point, the question arises as to whether AM detection in VFs can be usefully analysed with reference only to psychoacoustic research, essentially ignoring the fact

that it is a speech sound. VFs presented in isolation might not be perceived as a speech sound, but in a natural utterance in the context of a vowel environment this would not be the case.

Researchers have posited of a 'speech mode' of perception in which speech units are processed in a different way to other acoustic inputs. Commonly cited evidence in favour of this view comes from the phenomenon known as *sine-wave speech* in which a combination of sine waves with frequencies reflecting a changing set of speech formants typical of a certain utterance can either be heard as 'science-fiction' type noises or a comprehensible spoken phrase. However, once heard as speech, subjects are unable to switch back to hearing incomprehensible noises (Remez et al., 1981).

Whether or not speech entails a special form of perception, there is some suggestion in the literature that the perceptual challenges presented whilst processing speech influence listeners' performance in basic psychoacoustic tasks such as AM detection. For example, in an early pitch discrimination experiment, Flanagan and Saslow (1957) showed that discrimination thresholds for the f_0 of natural vowels are higher than for pure-tones of the same frequency. More recently, Cosgrove et al. (1989) studied the detection thresholds for upward and downward F2 transitions in synthesised vowels with four formants. Thresholds were significantly higher under all conditions than for a pure tone with f_0 equal to the starting F2 frequency in the vowels (1420 Hz). However, other basic psychoacoustic processes may be unaffected by speech. For example, Healy and Bacon (2006) concluded that the critical band (CB) resolution employed by listeners to decode running speech matches 'reasonably well' that obtained from traditional psychoacoustic measurement procedures.

2.2.4 Summary

Early work on AM detection for simple broadband stimuli suggests that the range of m tentatively proposed in 2.1 would be easily detectible by listeners, even though the specific acoustic characteristics of modulation within VFs have been shown to affect AM detection. For example, the short duration typical of VFs, as well as the variable spectral patterns of the fricative noise and the onset and offset patterns of modulation potentially caused by physiological factors involved in producing simultaneous voicing and frication could all riase AM detection thresholds.

Furthermore, there is ample evidence in the literature that the combination of a tone with the AM noise has a complex effect on detection which depends on a variety of acoustic parameters of both noise and tone. Direct evidence comes from three studies that introduced a periodic component to the noise in different ways. In all cases, the amplitude and phase of the introduced component in relation to the modulating

signal were key in determining whether detection was impaired or enhanced and to what degree. Other suggestions of interaction come from tone-on-noise masking and CMR, both of which may have an effect on the perceived levels of the tone or noise components.

AM detection could also be affected by the properties of the speech environment in which the VF is found; for example, nonsimultaneous masking by the preceding vowel is a possibility. Furthermore, the 'framing' of the VF in a speech context may interfere with the basic task of AM detection, although evidence is limited to a few studies. Finally, impressionistic evidence is taken from the improvements in synthesis quality and cohesion that have been achieved by modulating the noise component of mixed source speech sounds.

2.3 Speech Perception

Amplitude modulation of the noise component has generally not been considered as a possible cue to the voicing distinction in fricatives; in fact, only Strope and Alwan (2001) have previously studied the possibility. Despite this lack of attention, AM in noise, both in general and specifically in fricative sounds, has received some consideration from a perceptual viewpoint. Fant (1960) first noted that source-source interaction occurred as "periodic and synchronous" modulation of the frication source by phonation and early work by Stevens (1971) illustrated the detailed acoustic consequences of modulated turbulence noise. Psychoacoustic work on subjective judgements of noise quality identifies AM with a sensation of 'roughness' (Cox, 2008; Jeong, 1999; Zwicker and Fastl, 1999; Milosevic et al., 2004; Daniel and R.Weber, 1997; Terhardt, 1974). By introducing this percept into fricative noise in synthesised speech, researchers have achieved improvements in the quality of the output. AM also aids in the perceptual integration of noise and voicing sources (Hermes, 1991).

2.3.1 Amplitude Modulation Enhances Quality and Perceptual Integration of Sources

In simple models of VFs, the individual contributions from voicing and frication sources are summed to form the output: voicing as a volume-velocity source at the glottis and frication as a pressure source at the supraglottal constriction.

Flanagan's electrical analogue model was one of the first to incorporate AM of the fricative source (Flanagan and Cherry, 1969). Band-passed Gaussian noise (0.5–4 kHz) was multiplied by the squared volume velocity at the constriction exit, including the d.c. component.

Sondhi and Schroeter (1987) employed a similar model for aspiration at the glottis, gated by a threshold Reynolds number; for frication they placed a volume-velocity source 0.5 cm downstream of the constriction exit (or at the lips for /f, v, θ, ð/) to improve the subjective effect.

Klatt treated aspiration and frication identically, modulating the noise source by a square wave (50% burst duration) that was switched on during voicing, to achieve the effect he wanted, remarking that it is "not necessary to vary the degree of amplitude modulation ..., but only to ensure that it is present" (Klatt, 1980).

In Scully's work (Scully, 1990; Scully et al., 1992), noise generation was based on Stevens' static experiments (Stevens, 1971). Motivated by perceptual test results, aspiration noise was modulated using the rapidly-varying glottal area.

In the Portuguese articulatory synthesiser of Teixeira et al. (2005), the volume velocity at the fricative constriction is based on the flow at the glottis and transfer functions computed for noise sources at several instants during an f_0 pitch period, allowing them to activate and deactivate.

Although no empirical data exists in respect of VFs, sounds with temporal patterns or envelopes tend to be perceived as coming from the same 'auditory object' (Bregman, 1990). The result is a unified auditory object percept, rather than separate source sounds. Evidence comes from casual observation of the results of speech synthesis: recall that Hermes (1991) demonstrated this for speech sounds by improving the coherence of synthesised breathy vowels through AM of the aspiration noise. In this way, "noise was no longer perceived as a separate sound, but integrated perceptually with the strictly periodic part of the signal". Klatt (1980) also successfully used square-wave modulation of the noise component to enhance the quality of synthesised speech for both vowels and fricatives.

2.3.2 From Quality to Cue

> *"From the existing evidence it can indeed be concluded that, given the opportunity, listeners will make use of any cue for a given phonetic distinction. This general observation suggests that... the concept of cue has limited theoretical relevance. As a practical manner it is useful, even essential, in dealing with the acoustic basis of speech perception. But the sensitivity to the many and various cues for a phonetic segment suggests... that listeners are perceiving just what all the cues have in common — namely, some economical representation of the coherent process underlying the peripheral articulation."* Repp and Liberman (1990)

In reality, considering AM of noise from a psychoacoustic 'roughness' perspective, an 'auditory object' perspective, or a speech synthesis quality perspective may be too simplistic an approach.

Modern versions of articulatory-based theories of speech perception posit that all the acoustic consequences of an articulation, or 'cues', can be used by listeners to facilitate perception. Both the *Motor Theory (MT)* (Liberman et al., 1967; Liberman, 1996) and *Direct Realism Theory (DRT)* (Fowler, 1991, 1996) invoke a layer of articulatory representation between the acoustic signal and phonological and higher order processing. According to MT, listeners perceive nuero-sensory commands to the articulators, whilst DRT focuses on the gestures of the articulators themselves[6]. The existence of

[6] A detailed discussion of current theoretical treatments of speech perception will not be presented

an articulatory percept means that multilateral phonetic variation is easily explained: listeners have tacit knowledge of human articulatory abilities and limitations and are able to use this knowledge to map acoustic events to articulatory representations and thereafter to phonological categorisation, word recognition and general higher level linguistic processes.

Purely acoustic approaches to perception, such as the *Feature Detector* hypothesis associated with Stevens and Blumstein (see Stevens and Blumstein (1978); Blumstein et al. (1977); Blumstein and Stevens (1979, 1980, 1981)), contrast with this in positing a direct mapping from the acoustic signal to identification of phonemes, without the need for an articulatory reference layer, i.e., there is enough information contained within the acoustic signal to permit retrieval of the underlying phonological code. Acoustic invariants in the signal are said to correspond to primary cues to which humans are biologically tuned, and whilst phonetic variability (e.g., speaker or context related) is not denied, the acoustic consequences of such variability, along with other, 'minor', invariant cues, are considered secondary.

Regardless of theoretical standpoint, the evidence that redundancy in phonetic 'cues' is actively taken advantage of by listeners warrants further attention. This redundancy has been demonstrated in the laboratory in many studies using phonetic cue-trading experiments: equal probability of a particular phonological categorisation can be attained with different combinations of 'primary' and (possibly multiple) 'secondary' cue settings, indicating perceptual equivalence (a *trading relation*) for these different combinations of multiple, possibly acoustically disparate cues.

Some examples of trading-relations for fricative sounds from the early cue-trading literature[7]: Repp et al. (1978) showed trading relations for pre-word silence duration, frication noise duration and rate of articulation of the sentence context in cueing the fricative/affricate distinction in word-initial position. Mann and Repp (1980) traded spectral shape of the frication noise and voiced formant transitions at frication offset for prevocalic /ʃ/ vs /s/. Bailey and Summerfield (1980) investigated a wide variety of cues to place and manner specification in stop consonants in medial position after [s]. For place, they showed perceptual sensitivity to the spectrum of the fricative at offset, the duration of the silent closure interval, the spectral relationship between the frequency of the stop release burst and following formants, and the spectral and temporal characteristics of the first formant transition. For manner, cues included the duration

here as the motivation of this work is not in providing evidence in support of, or against, any particular theory. Theories are mentioned here as part of the general introduction to cue-integration, cue-trading and cue-redundancy. For a comprehensive review of the current theoretical landscape, see Diehl et al. (2004)

[7]For a review of a wider range of findings from cue-trading experiments, see Repp (1982)

of silent closure, the frequency of the first formant at release, magnitude of the first formant transition, and the proximity of the second and third formants at release.

In the next section we consider, in detail, specific cues to the voicing distinction in fricatives that have been previously reported.

2.3.3 Potential Cues to the Voicing Distinction in Fricatives and their Trading Relations

A number of acoustic studies have measured fricative durations in a variety of phonetic conditions. Table 2.1 is a summary of these studies, listing mean fricative durations for voiceless and VFs in milliseconds.

The voicing distinction is strongly reflected in the reported durations: VFs are shorter than voiceless, although the mean magnitude of the difference varies from study to study, ranging from 27 ms (22%) in Baum and Blumstein (1987) to 69 ms (67%) in Jongman et al. (2000). Borzone de Manrique and Massone (1981) cite a larger difference in stressed (125%) than in non-stressed (83%) syllables; Baum and Blumstein (1987) found smaller differences in running speech than in single word citations, in agreement with the results from 'fast' and 'slow' groups in Crystal and House (1982). Across all the studies using English fricatives, there is less variability in the *absolute* differences in milliseconds (+54 ms, +30 ms, +69 ms and +27 ms) than the percentage change from voiced to voiceless (+106%, +42%, +67% and +22%). This may suggest that voiceless fricatives are generally longer by a set amount rather than a set proportion.

In the studies cited, there is disagreement regarding duration distributions for voiced and voiceless fricatives, and hence the efficacy of the acoustic measure as a potential cue. Borzone de Manrique and Massone (1981) suggest there is "little or no overlap between duration ranges", whereas Baum and Blumstein (1987) suggest that "while there is a tendency toward a bimodal distribution, there is considerable overlap between the two duration distributions for all fricatives and for all speakers"; they also report particular vowel contexts for particular speakers where VFs are longer than their voiceless counterparts.

Frication duration as a cue has also been demonstrated in perceptual studies. Denes (1955) played subjects unvoiced frication of duration varying from 50 to 250 ms spliced to a preceding environment [ju], where the duration of the vowel ranged from 50 to 200 ms. Listeners identified each token as either "(the) use" ([jus]) or "(to) use" ([juz]). As frication noise is lengthened, the percentage of voiced responses by subjects falls with all preceding vowel durations. At the shortest frication duration (50ms), voiced responses range from 75 to 100%, depending on preceding vowel length. At the

Study	[f]	[θ]	[s]	[ʃ]	Mean	[v]	[ð]	[z]	[ʒ]	Mean
Crystal and House (1982): Fast Speakers	-	-	-	-	94.5	-	-	-	-	47.1
Crystal and House (1982): Slow Speakers	-	-	-	-	114.3	-	-	-	-	54.1
Stevens et al. (1992)	94	-	108	-	101	64	-	78	-	71
Mair and Shadle (1996)	-	-	295	293	294	-	-	258	248	253
Jongman et al. (2000)	166	163	178	178	171	80	88	118	123	102
Baum and Blumstein (1987)	149	134	174	-	152	116	107	152	-	125
Pirello et al. (1997)	181	-	194	-	188	117	-	146	-	132
Borzone de Manrique and Massone (1981)	170	-	168	190	176	-	81	-	124	103

Table 2.1: Mean fricative durations in milliseconds, as reported by various studies. (Crystal and House, 1982, 1988) measured fricative durations for tokens occurring in naturally read texts for 'slow' and 'fast' readers. Both are presented in the table. Data reported by Mair and Shadle (1996) is for French and is averaged over three vowel environments. The data from Borzone de Manrique and Massone (1981) is for Spanish and is averaged over stressed and unstressed syllables.

longest frication duration (250ms), voiced responses range from approximately 0% with a 100 ms vowel to 30% with a 50 ms vowel. Although there is evidence of elementary cue trading of duration with preceding vowel length, no phoneme boundary or categorical effect (see Pickett (1999)) is evidenced.

In a more recent experiment, Cole and Cooper (1975) found a similar pattern with a clear phoneme boundary effect for both fricatives and affricates using a forced-choice paradigm with a 6-stimuli series where frication duration was shortened incrementally from its natural value. Figure 2.6 shows identification functions from their experiment with a sharp crossover between voicing category responses around stimulus value 4. There does not, however, appear to be any significant cue trading with preceding vowel

duration.

In the data from Stevens et al. (1992), investigating multiple perceptual cues to voicing, a three-way ANOVA performed on the results revealed a main effect showing more voiced responses when the frication duration was shorter ($p < 0.001$).

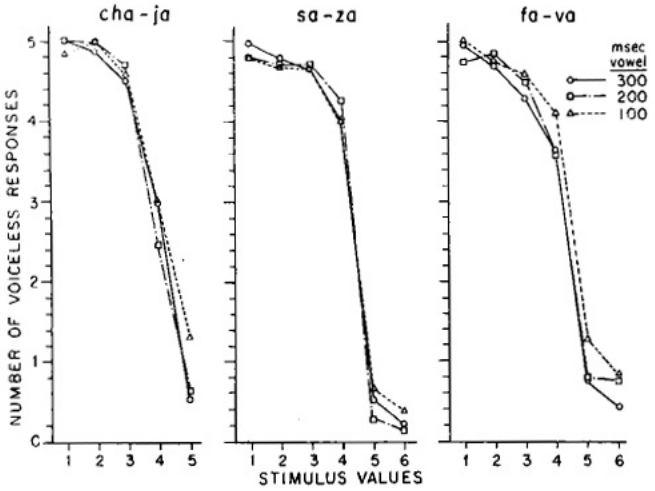

Figure 2.6: Identification functions for experimental series; from left to right: [tʃ]–[dʒ],[s]–[z] and [f]–[v]. Stimulus values from 1 (original length) decrease in steps of 1/6 of original length. Source: Cole and Cooper (1975).

Other perceptual data suggests that fricative duration does not cue the voicing distinction. Jongman (1989) investigated the minimum duration of frication noise required for correct identification of fricative consonants. He concluded that "subjects do not have a tendency to identify more fricatives as voiced as the frication duration decreases. Identification of the voicing category is quite good, independent of the duration of the frication noise, and frication noise duration *per se* can, therefore, not be considered a sufficient cue in the distinction among the voiced and voiceless fricatives."

In an experiment varying preceding vowel duration, offset 'structure' (described as the final amplitude contour of the vowel at the fricative boundary) and frication duration, Soli (1982) also found that duration was insufficient to cue the voicing distinction. Instead, vowel duration and cues in its offset structure[8] were primary and appeared to display trading relations. Soli argues, furthermore, that duration is only found to be an effective cue when unrealistic stimuli are used:

> The previous research on postvocalic fricative voicing has, in fact, employed somewhat stylised synthetic vowels. For example, Denes (1955) used two-formant synthetic vowels, while Derr and Massaro (1980) used full, five-formant vowels but with linearly approximated formant transitions.

[8] Although importantly, 'offset structure' as defined and tested at first proved not to be the important cue in the vowel structure at the boundary with the fricative. Acoustic analyses following the perceptual testing revealed that the important cue in vowel offset was the proportional duration of the formant transitions.

Stevens et al. (1992) investigated the effect of offset and onset transitions from the vowel by varying the extent of the first formant transition slopes. Stimuli were synthesised with 'full', 'short' and 'no' transitions. Results show a trading relation between duration of frication noise and duration of formant transitions: short (70 ms) stimuli with full transitions were sufficient to elicit the majority of voiced responses. They conclude that reducing the extent of the transitions is perceptually equivalent to increasing the duration of frication during which there is no glottal vibration.

The concept of duration of source overlaps was developed by the same authors. A Klatt synthesiser was used to generate /VFV/ tokens with glottal excitation extending between 10 and 40 ms into the frication noise at either the /VF/ boundary or the /FV/ boundary. Amplitude of the glottal vibration relative to the adjacent vowel varied from 0 dB to -15 dB. More voiced responses were attested when duration of glottal excitation was longer ($p<0.001$) and more glottal excitation was needed at the /FV/ than at the /VF/ boundary to elicit the same probability of a voiced response. All variables appear to enter into trading relations with each other: 50% voiced responses could be elicited with 20 ms of 0 dB glottal vibration or 32 ms of -10 dB glottal vibration, both at the /VF/ boundary; at the /FV/ boundary, ∼10 ms more glottal excitation is needed for an equivalent response.

In an acoustic experiment, Pirello et al. (1997) also found that glottal excitation at the fricative boundaries was key to voicing. Using a classification metric of a minimum of 30 ms of contiguous 0–10 dB voicing at either onset or offset of frication for a 'voiced' classification, 93% of tokens were classified correctly.

These results of Stevens et al. (1992) were used to estimate the duration of the *interval in which there is no glottal vibration* that is necessary to elicit perceptual responses of 50% voiceless. For the most realistic stimuli, with glottal vibration at the /VF/ boundary and natural glottal amplitudes, it is concluded that about 60 ms of frication without glottal vibration is sufficient to elicit a voiceless response. In a further experiment, it was found that a stimulus with 65 ms of no glottal vibration at the /VF/ boundary elicited only 33% 'voiceless' responses in comparison to the 50+% in the previous experiment. They conclude that, with slight perceptual adjustment based on F1 transitions, listeners are likely to use "length of frication without glottal excitation" as a cue to the voicing distinction and that the threshold value is approximately 60 ms. Re-examination of the acoustic results showed that all voiceless fricatives and approximately 90% of VFs satisfied this criterion.

Massaro and Cohen (1976) examined cues to the voicing distinction with word-initial /zi/ and /si/ tokens. Fricatives of 130 ms in duration were synthesised with voice onset time (VOT) varying from 70 ms to 130 ms (completely voiceless) in four steps. Listeners were asked to judge stimuli on a 4-step scale from /z/ to /s/. As VOT was extended,

more voiceless-like judgements were attested. With an *initial* voiceless period of 60 ms (70 ms VOT), stimuli were judged as being more voiceless-like, but not completely voiceless. Fully voiceless responses are not elicited until the initial voiceless period is equal to the full length of the fricative, suggesting that voiceless period may not be an independent perceptual cue in word-initial position.

Finally, Strope and Alwan (1998) have made initial attempts to characterise the role of AM of frication noise, the focus of this thesis, in cueing the voicing distinction. In a perceptual experiment, the subject's task was to correctly identify [s] or [z] where tokens had been high-pass filtered at 3 kHz, which the authors suggest leaves only AM of the frication noise as a possible cue. Tokens were mixed with wide band noise of increasing SNR to the frication noise, progressively decreasing the effective modulation depth of the AM in [z] available to distinguish it from the unmodulated [s]. Using a 2AFC adaptive procedure, the 79.4% threshold was estimated at 0 dB SNR (wideband masker to fricative noise). Given that naturally spoken [z] tokens were used, and thus the starting m value unknown, it is not possible to translate this SNR value into an effective modulation depth.

2.3.4 Summary

AM was recognised as an acoustic feature of voiced frication in the early speech literature. Work on speech synthesis has led researchers to include simple and complex simulations of AM into their frication generation models in order to better replicate natural VFs sounds, and have achieved higher quality, more cohesive sounding synthesis as a result.

AM may be more important than simply enhancing quality though. Modern theories of speech perception emphasise listeners' abilities to integrate a wide variety of seemingly disparate 'cues' in solving the speech perception problem. In fact, multiplicity and redundancy of cues is necessary and desirable in most circumstances of sub-optimal speech transmission.

The literature on cue-trading in phonetic perception shows how ubiquitous this redundancy is. More specifically, studies investigating the voicing distinction in fricatives, the direct object of interest here, have uncovered a large number of cues that are disparate in time, frequency and in the case of AM, even in domain (spectral vs envelope): turbulence noise and periodical glottal sources are perceived semi-independently and in respect of their overlap durations and relative amplitudes at multiple locations (onset, middle and offset) throughout the fricative in addition to cues from the duration, amplitude and formant transitions of surrounding vowels. A number of these cues were investigated in relation to AM in Chapter 5

Importantly, a majority of these studies confirm that cues are traded on multiple dimensions and listeners can probably take advantage of most, if not all, in perceiving any one token. AM has been shown to be perceptible and effective in cueing the voicing distinction when all other cues are neutralised. How, though, do listeners integrate the AM percept with other 'primary' cues to form a unified percept of VFs?

Chapter 3

Acoustic Study

3.1 Introduction

The research presented in this section aims to address the need for more extensive data on the acoustic characteristics of AM in voiced frication and fricatives.

Recall from Equation 1.1 that modulation depth can be specified or measured at the fundamental frequency, f_0, as well as harmonics thereof (i.e., $h \in 2..H$). Casual observation of modulation spectra confirms that AM at harmonics above f_0 are weak; in fact, it is unlikely that modulation could be measured above natural background fluctuation for anything above the second harmonic, i.e., m_3. This may be a result of the AM generation mechanism, which concentrates modulation at the fundamental. It was thus decided to restrict measurement of m to the fundamental plus first two harmonics.

Recall further that in the case of AM noise accompanied by a period component (pure tone or complex), the phase relationship, ϕ, of the modulating signal, $a(n)$, to the fundamental of the periodic component is also a measurable acoustic variable. Given the limitations of time, it was decided to focus exclusively on the measurement of modulation depth and its correlation to a number of variables intrinsic to VFs: a measure of voicing strength, v, fundamental frequency, f_0, and fricative place of articulation (i.e. fricative noise spectrum). This decision was motivated by the primary role assigned to modulation depth in the perceptual literature. In response to the findings of the psychoacoustic experiments performed subsequent to the acoustic study and in evaluating the output of this research, it will be argued that acoustic measurement of phase is an essential direction for further work in this area.

3.2 Method

3.2.1 Speech recordings

Two sets of fricative recordings were designed to capture the range of aeroacoustic conditions achievable by the human vocal apparatus in steady phonation and those typically realised in fluent speech. The *sustained fricative* set included laryngeal measurement of the vocal fold vibration and calibrated sound pressure from a microphone at a fixed distance from the subjects' lips for 16 subjects (12M, 4F). The *fluent-speech fricative* set provided a more natural environment with nonsense words framed in a standard phrase for 8 subjects (4M, 4F). Four subjects took part in both experiments. Numbers of male and female subjects reflected availability, while providing at least four of each sex, so gender effects are treated cautiously. Number of repetitions per speaker was chosen for each corpus by balancing the need for statistical significance against the need to keep recording time for each speaker as short as possible (maximum 30 minutes) to avoid loss of concentration or degradation of voice quality.

Subjects were unpaid staff and students of the University of Surrey, age range 20–35, all with British RP accents. Before recording, it was verified that subjects were able to produce all vowels and fricatives correctly and that they were aware of the voiced/voiceless distinction. The distinction between [θ] and [ð] tended to cause difficulty, as well as the voiced postalveolar fricative [ʒ].

Sustained fricatives

Fricatives [ʒ,z,ð,v] were spoken in isolation. Both male and female subjects was asked to produce two types of utterance at three pitch settings, $f_0 \in \{125,150,175\,\mathrm{Hz}\}$. Having three settings enabled an analysis for f_0. Each recording was preceded by a pitch-reference tone played through a loudspeaker and short (2-s) pause to allow subjects to tune their pitch.

The first utterance type was an uninterrupted fricative where the subject smoothly adjusted loudness from their quietest possible fricative to loudest, and again to quiet, and loud ($\sim$3 s in total). This utterance type was designed primarily to provide a continuum of voicing strength/modulation data but subjects sometimes struggled to maintain the aerodynamic vocal tract conditions necessary for successful completion of the 3s sequence.

A second utterance type was conceived, consisted of three separate sustained fricatives of increasing intensity ($\sim$1 s each). This design allowed subjects to breath between each

fricative and thus more easily produce a section of smooth voiced frication with stable pitch.

In total 24 recordings were made for each speaker (4 fricatives $\times$ 3 f_0 values $\times$ 2 types).

Speech audio and electroglottograph (EGG) signals were captured simultaneously on PC by a Creative Labs Audigy sound-card via a Sony SRP-V110 desk (2 channels at 44.1 kHz, 16 bit): mono audio from a Beyerdynamic M59 microphone, and EGG from a Laryngograph Lx Proc PCLX with adult-sized electrodes. The microphone was calibrated by comparing a 1 kHz tone played through a loudspeaker at 10 cm to an SPL measurement made with a Brüel and Kjær Type 2240 SPL meter at the same distance. The microphone was calibrated once using a 1 kHz tone played through a loudspeaker at a constant gain setting.

During recording, subjects placed their head in a movement-restricting support and were instructed to keep still in order to control the lip-microphone distance to within a few millimetres of 10 cm, at lip level and $\sim$45° to the line of sight.[1] The EGG signal provided accurate pitch information which was used by the modulation depth estimation algorithm.

Fluent fricatives

Tokens of F=/ʃ,ʒ,s,z,θ,ð,f,v/ were recorded in nonsense /VFə/ words with V=/ɑ,i,u/, embedded in the phrase "What does /VFə/ mean?", using an acoustically sheltered cubicle and equipment as above. The use of carrier sentences rather than words only was to encourage natural, flowing speech. Subjects were given two kinds of prompt. A randomised, printed list of sentences (see Appendix B) was provided. Nonsense words were spelled using normal English orthographical rules and warnings printed next to sounds causing difficulty as follows:

as in 'think' to denote /θ/
as in 'the' to denote /ð/
as in 'beige' to denote /ʒ/

The printed list was supplemented with an audio recording played through single-ear headphones, allowing the subject to hear their own utterances at the same time as the audio prompt.[2]. The sentences, as presented on the visual handout, were recorded by

[1] A short lip-microphone distance helped to capture quiet frication over any background or electric noise.

[2] A pilot experiment revealed that some subjects had difficulty keeping their place on the printed list while speaking. The audio prompting was designed to aid them, but also as a natural control on speech rate and intonation.

the author in blocks of ten to the rhythm of a metronome whose beats were spaced by 1.5 s. Sentences began immediately after a beat and were spoken at a rate such that the end of the sentence occurred just prior to the next beat. The gap until the next beat was then left blank for the subject to repeat the sentence. A small break occurred after every ten sentences. To promote natural, fluent speech, subjects were left free to move their head.

For each speaker, 216 sentences were recorded (9 reps. of every /VF/ pair). Furthermore, since durational measurements were to be made by hand, more tokens would require more time spent on manual annotation. Nine repetitions per speaker translated to approximately 30 minutes of recording time, which proved to be both manageable for speakers and for manual segmentation. Furthermore, 9 repetitions per speaker provided a good number of tokens for statistical analysis.

3.2.2 Preparation of Recordings for Acoustic Measurements

Before application of the preprocessing and measurement algorithms, the recorded waveforms were manually prepared for measurement. For sustained fricatives, silent intervals were cropped from the recording in the case of the second utterance type. For fluent-speech fricatives, onset and offset points of fricatives embedded within the nonsense words and carrier sentences were manually identified using graphical waveform software, as exemplified in Figures 3.1 to 3.3.

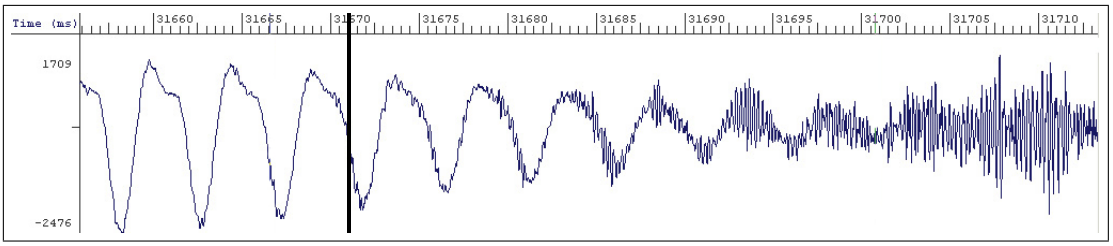

Figure 3.1: Example of manual segmentation annotations (frication onset, F_{ON}, left marker) applied to /VF/ boundary of [uʒə] spoken by HC.

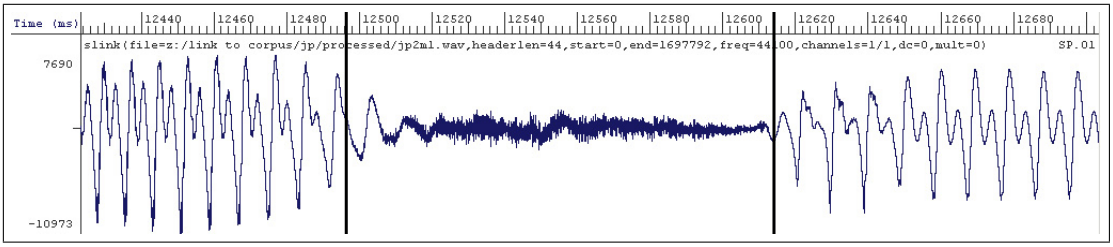

Figure 3.2: Example of manual segmentation annotations (frication onset, F_{ON}, left marker; frication offset, F_{OFF}, right marker) applied to /VF/ boundary of [ɑsə] spoken by JP.

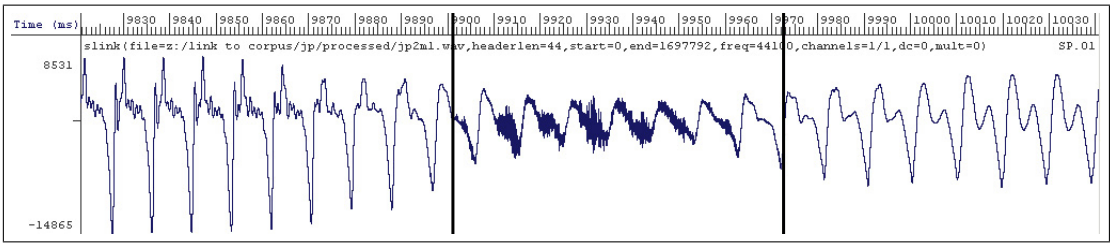

Figure 3.3: Example of manual segmentation annotations (frication onset, F_{ON}, left marker; frication offset, F_{OFF}, right marker) applied to /FV/ boundary of [ɑzə] spoken by JP.

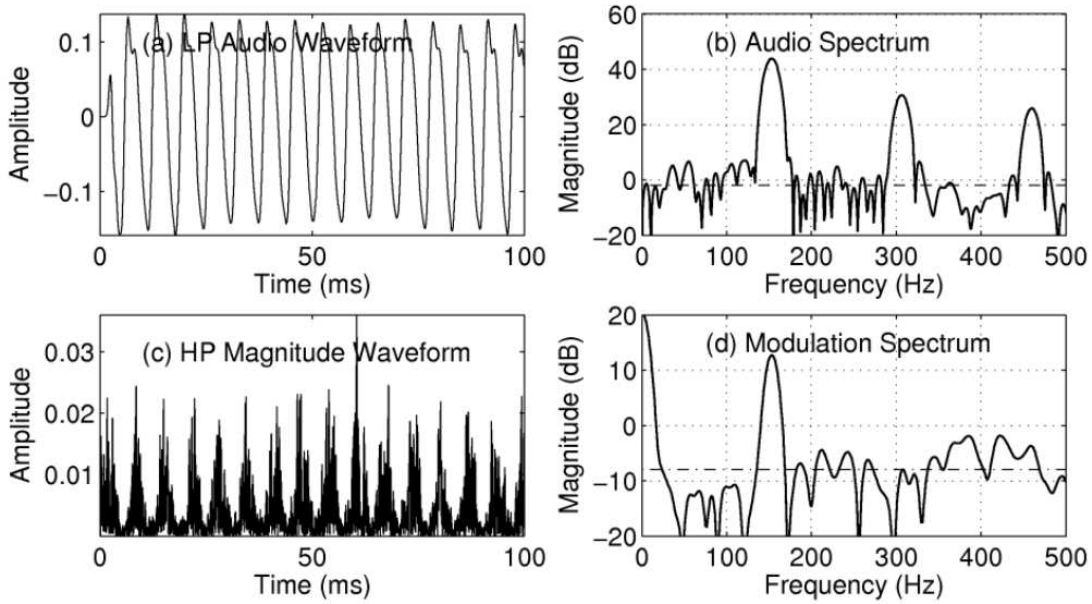

Figure 3.4: Illustration of the harmonic structure of the voicing signal (top row) and the modulating signal (bottom row) for 100 ms of [z] ($f_0 \approx 150$ Hz). (a) Audio waveform low-pass filtered at 1 kHz. (b) Audio spectrum up to 500 Hz. (c) Magnitude of waveform high-pass filtered at 9 kHz. (d) Modulation spectrum. Dashed lines in spectra indicate noise floor.

3.2.3 Measuring modulation depth

Estimating the modulation index m_h

With modulated broadband noise, the carrier signal $w(n)$ is an unknown random variable, which can be modelled as Gaussian white noise, and the signal $x(n)$ is as in Eq. 1.1. To estimate m_h, the instantaneous magnitude of the signal is taken $|x(n)| = |w(n)|a(n)$ which, unlike the modulated noise signal, contains a periodic component at f_0 and its strength is proportional to m_1. To extract this component, the Fourier transform is computed $\bar{X}(k) = \mathcal{F}\{|x(n)|\}$, applying a Hamming window and zero padding:

$$\bar{X}(k) = \mathcal{F}\{|w(n)|\} \otimes \left[\Delta(k) + \sum_{h=1}^{H} \frac{m_h}{2} \left(\Delta(k - hk_0)\, e^{j\phi_h} \right) + \sum_{h=1}^{H} \frac{m_h}{2} \left(\Delta(k + hk_0)\, e^{-j\phi_h} \right) \right], \tag{3.1}$$

where $\otimes$ denotes convolution, $\Delta(\cdot)$ the Fourier transform of the window function, and $k_h = hNf_0/f_S$ is the frequency bin that contains harmonic hf_0. Figure 3.4(d)

shows the *modulation spectrum*, $\bar{X}(k)$, for frication noise from a [z] token modulated at $f_0 \approx 150\,\mathrm{Hz}$, where the spike occurs. Hence, modulation index m_h can be estimated by comparing the magnitudes of the Fourier coefficients at hf_0 and d.c.: $\hat{m}_h = 2\left|\bar{X}\left(k_h\right)\right| / \left|\bar{X}\left(0\right)\right|$. where the factor of two leads to an estimate of the standard modulation index.

Allowing for pitch variation

Although the processing window is short enough to exclude major changes in fundamental frequency, pitch variation within a window smears the modulation energy at each harmonic. To compensate for variable pitch and spectral smearing from windowing, our estimate $\hat{m}_h$ was based on the area under the spike at k_h and above the noise floor, including adjacent bins as appropriate [3]. This defined $\tilde{k}_h$ as the contiguous set of bins under the k_h spike (see Fig. 3.4(d)). For the spike around zero frequency, a range of bins was also aggregated, $\tilde{0}$. Thus, with noise floor $\hat{\theta}^2 = \frac{1}{N}\left(1 - \frac{2}{\pi}\right)\sum_{k=0}^{N-1}|X(k)|^2$, a estimate similar to that above was formed:

$$\hat{m}_h = 2\left(\frac{\sum_{\tilde{k}_h}\left|\bar{X}(k)\right|^2 - \hat{\theta}^2}{\sum_{\tilde{0}}\left|\bar{X}(k)\right|^2 - \hat{\theta}^2}\right)^{1/2}. \tag{3.2}$$

[3] The present method diverges here from that used in Pincas and Jackson (2004).

3.3 Application to Speech

3.3.1 Periodic energy mixed with noise

Since VFs comprise periodic energy mixed with frication in the time waveform and much of the spectrum, it is not trivial to isolate the noise component for analysis. The f_0 component itself is confined to low frequencies (< 400 Hz) and can easily be removed by high-pass (HP) filtering without losing any significant amount of fricative noise. The choice of HP cut-off frequency has some effect on estimates of m, however, and so simulations were required before deciding on a final value. The results of these simulations are discussed in Section 3.3.4.

Above f_0, bands of periodic energy, or voicing harmonics, persist into the higher spectral regions of the fricative noise, especially near formant frequencies.

For most speech sounds, interest is focused on the first two or three formants (up to perhaps 4 kHz) as higher formants tend to be weaker and are less important perceptually. In normal, fluent VF speech, voicing is often weak and its formants are rarely detectable much above 3 kHz; in strong fricatives with a loud voicing component (as in our sustained fricatives corpus), formants can be found up to 5 or 6 kHz.

Consider the spectrum of a strongly voiced [v] in Figure 3.5. Fig. 3.5(b) shows the dominant periodic energy in the 0–4 kHz region (a harmonic spectrum with four defined formant peaks at 1.3, 2.2, 3.2 and 3.7 kHz); in Fig. 3.5(c), the spectrum is purely aperiodic, with no harmonics that can be ascertained in the 7–16 kHz range; in Fig. 3.5(a), 4–7 kHz contains mainly aperiodic energy though with a defined formant at 6.2 kHz, which can be seen in both the spectrum and the spectrogram in Fig. 3.5(d).

The effect on apparent modulation depth of mixing periodic energy with frication noise should be considered. Given that formants are damped resonances excited periodically by voicing at f_0, they will tend to have a fluctuating envelope similar to that of the aperiodic component. Unless the peaks are in phase with the bursts of frication, the presence of voicing will attenuate the apparent modulation depth of the noise. Consider the fricative [v] in Figure 3.5. The spectrogram shows strongly modulated frication noise above 4 kHz, as well as fluctuating peaks in formant energy at lower frequencies. Careful inspection reveals that the pulses of frication are out of phase with the pulsed formant energy. Amplitude envelopes (or modulation signals, $a(n)$) for different frequency bands are shown underneath as Figs. 3.5(e) and (f), showing how they differ in phase. Fig. 3.5(e) compares amplitude fluctuation in the overwhelmingly periodic, 1–4 kHz band (thick line, cf. Fig. 3.5(b)), to the mainly aperiodic, 7–16 kHz band (thin line, cf. Fig. 3.5(c)). The phase difference between the two modulation signals is $\sim 170°$.

Envelopes in Fig. 3.5(f) demonstrate the attenuation of apparent modulation from combining the out-of-phase periodic and aperiodic energy components. HP filtering with cut-off $f_{\mathrm{HP}} = 1\,\mathrm{kHz}$ (thick solid line) removes the f_0 component *and first formant* but leaves the remaining periodic and aperiodic energy intact. Since the voicing source is stronger, the modulation signal is dominated by the periodic formant energy; the similarity in phase to the 'periodic only' 1–4 kHz band (thick solid line in Fig. 3.5(e)) confirms this. However, in comparison to the 'periodic only' case, the depth of modulation has been reduced; this is due to the out-of-phase aperiodic energy in the region 4–16 kHz. Raising f_{HP} to 3.5 kHz (dashed line) excludes most (but not all) of the periodic energy (see formant at 3.7 kHz in the spectrum and spectrogram), which evens out the periodic and aperiodic components. Modulation shape and depth are disrupted, and the phase of the modulating signal resembles neither of the previous cases. A further increment to $f_{\mathrm{HP}} = 4\,\mathrm{kHz}$ (thin line) excludes the last strong formant (a weaker one remains at ∼6 kHz) and the resulting envelope is similar, if weaker, to the 'aperiodic only' 7–16 kHz band (thin line in Fig. 3.5(e)).

Since we are interested only in modulation of the frication noise, it is paramount that the *aperiodic component* is successfully isolated before applying Eq. 3.2 to estimate the modulation depth. As Figure 3.5 demonstrates, failure to remove periodic energy can seriously affect the accuracy of m_1 estimation for the frication noise. Periodic components could be removed by HP filtering with f_{HP} high enough to exclude all likely periodic energy.

Although fixing f_{HP} at a higher value has the advantage of effective removal of periodic energy, it substantially limits the bandwidth of noise from which modulation depth is measured. This causes two problems: first, modulation is unlikely to be uniform throughout the frication noise spectrum (see Sec. 3.3.2); second, filtering AM noise removes some modulated sidebands which gives under-estimated modulation depth (see Sec. 3.3.5).

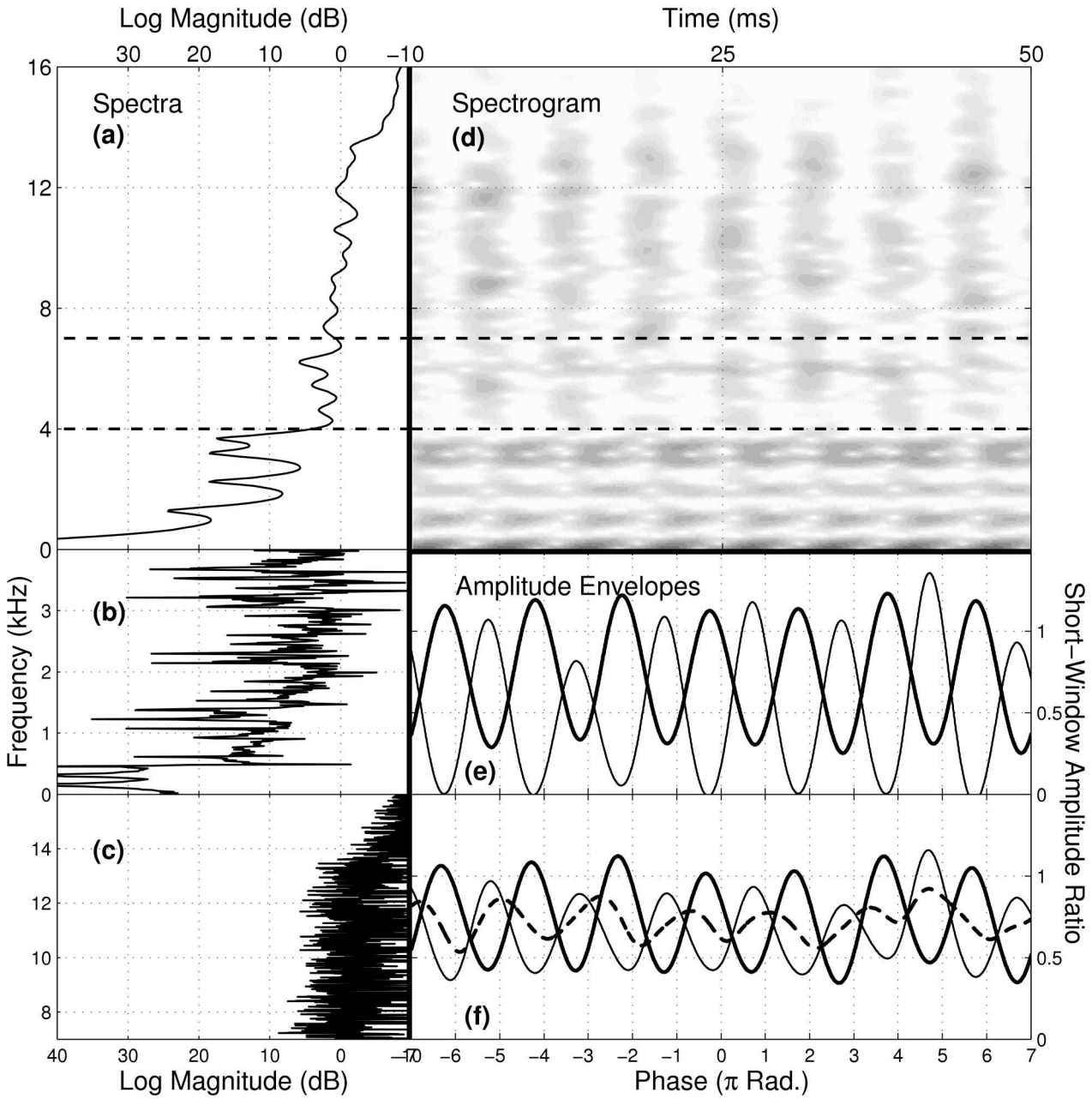

Figure 3.5: (a) LPC spectrum (order 40), (b) Close-up of spectrum in region 0–4 kHz, (c) Close-up of spectrum in region 7–16 kHz, (d) Spectrogram (5 ms, Hanning window, 4× zero-padded, fixed grayscale, frequency-aligned with LPC spectrum and time-aligned with amplitude envelopes), and (e,f) Amplitude envelopes (magnitude signal, low-pass filtered at 200 Hz) for 50 ms section of sustained [v] ($f_0 \approx 153$ Hz, f_s = 32 kHz). Individual amplitude envelopes are for different frequency bands, f_{BP}. (e) $1 \leq f_{BP} \leq 4$ kHz (thick line, periodic energy) and $7 \leq f_{BP} \leq 16$ kHz (thin line, aperiodic energy); dashed horizontal lines on spectrogram identify these frequency regions. (f) $1 \leq f_{BP} \leq 16$ kHz (thick line, mainly periodic), $3.5 \leq f_{BP} \leq 16$ kHz (dashed line, balanced mix of periodic and aperiodic) and $4 \leq f_{BP} \leq 16$ kHz (thin line, mainly aperiodic).

3.3.2 Non-uniformly modulated noise

Thus far, the noise signal has been treated as Gaussian white noise. In VFs, the carrier noise $w(n)$ is not white, but coloured (filtered) depending on place of articulation. The spectral composition of the noise does not directly affect the modulation of different frequency regions. However, it cannot be assumed that the mechanism responsible for modulation in fricatives produces uniform modulation across all frequencies; in fact, spectrograms of VFs suggest that noise in very high frequency regions (>8 kHz) is more modulated than in the main region (3–7 kHz). More work is needed to understand how the modulation mechanism produces uneven modulation depths across the noise spectrum.

Figure 3.6 shows a short portion (100 ms) of a strongly modulated [ʒ] that happens to lack strong voicing formants, allowing analysis of different frequency bands without interference from periodic energy. In the spectrogram, the frication noise looks modulated throughout the spectral range, but the weaker noise above 10 kHz comes in more distinct and separated bursts compared to the mid-range noise. This observation is borne out by analysis: amplitude envelopes for three spectral bands (magnitude signals, low-pass filtered at 700 Hz to catch the first few modulation harmonics) illustrate variations in the modulation signal through the noise spectrum. In the 3–6 kHz range (Fig. 3.6(b)), the modulation signal is noisy and its fundamental is weak ($m_1 = 0.56$). For 6–10 kHz (Fig. 3.6(c)), m_1 grows to 0.71, the waveform becomes more regular, and the periodic structure of the modulation signal emerges, with steep-sided, rather than sinusoidal, pulses. At 12–22 kHz (Fig. 3.6(d)), modulation at the fundamental is almost complete ($m_1 = 0.98$) and the waveform has regularised into a train of sharp (steep-sided) pulses separated by a noticeable gap. This is akin to the 'fundamental saturating under the action of its harmonic' described by Crow and Champagne (1971), where the fundamental can increase no further; instead, a significant harmonic structure develops where the modulation signal begins to adapt from sinusoid to pulse train. Thus, basing measurement of noise on the upper frequency bands could lead to an over-estimation of m_1 with regard to the full spectrum of frication noise. To balance the need for effective removal of periodic components and accurate estimation of modulation depth, the VF signals were pre-processed using a technique designed to segregate periodic and aperiodic energy as well as high-pass filtering to fully remove the low-frequency voicing component.

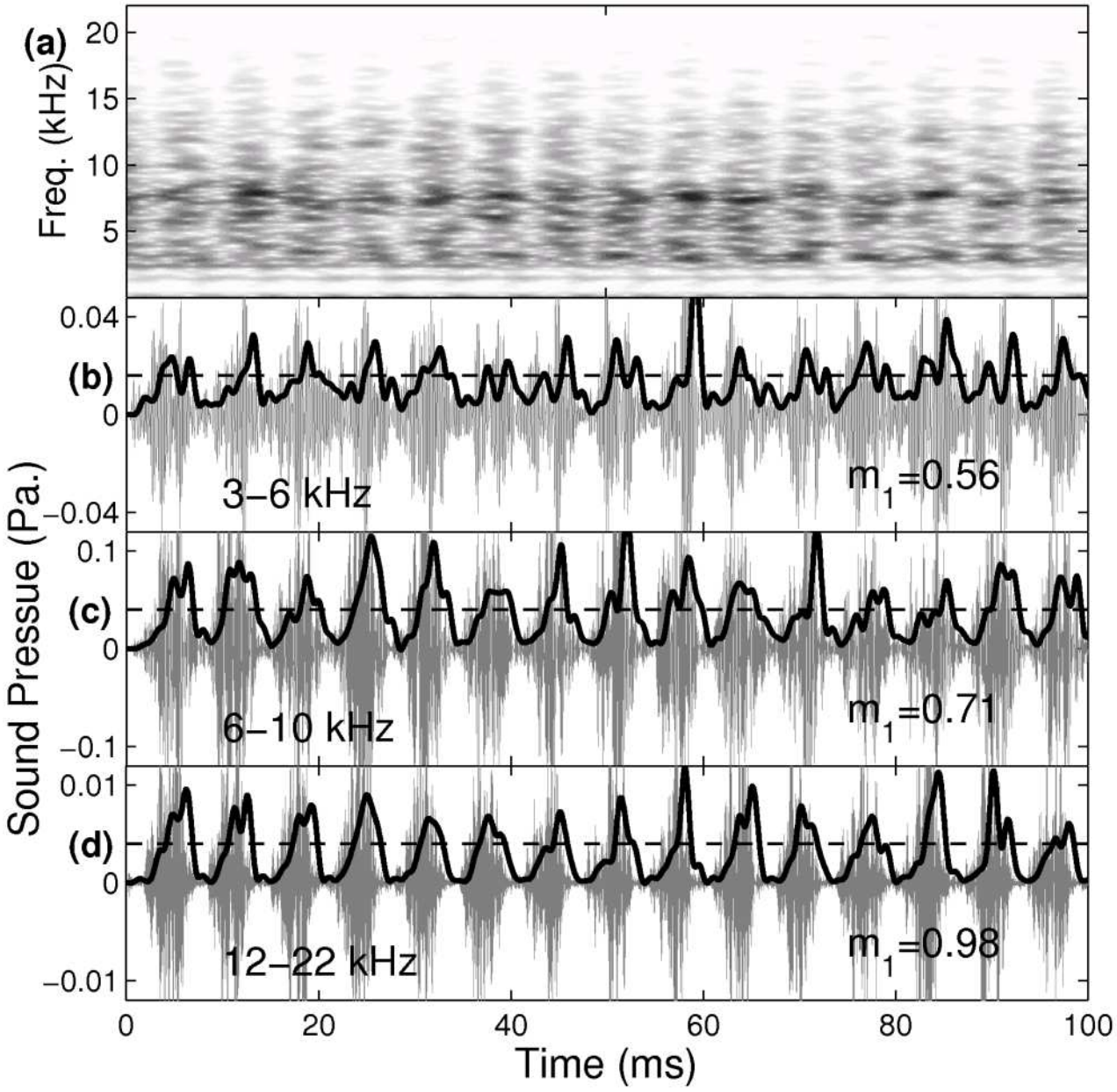

Figure 3.6: (a) Spectrogram, and (b,c,d) Time-aligned waveforms (light grey) with amplitude envelopes (black lines, magnitude signal low-pass filtered at 700 Hz) for 100 ms section of sustained [ʒ] ($f_0 \approx 152$ Hz, $f_S = 44.1$ kHz). Individual amplitude envelopes are for different frequency bands, f_{BP}, with axes scaled to $\pm 2\times$ RMS amplitude (indicated by dashed lines; notice the different scale for each band). (a) $3 \leq f_{\mathrm{BP}} \leq 6$ kHz; (b) $6 \leq f_{\mathrm{BP}} \leq 10$ kHz; (c) $12 \leq f_{\mathrm{BP}} \leq 22$ kHz. $\hat{m}_1$ values estimated for individual frequency bands as in Sec. II.B.

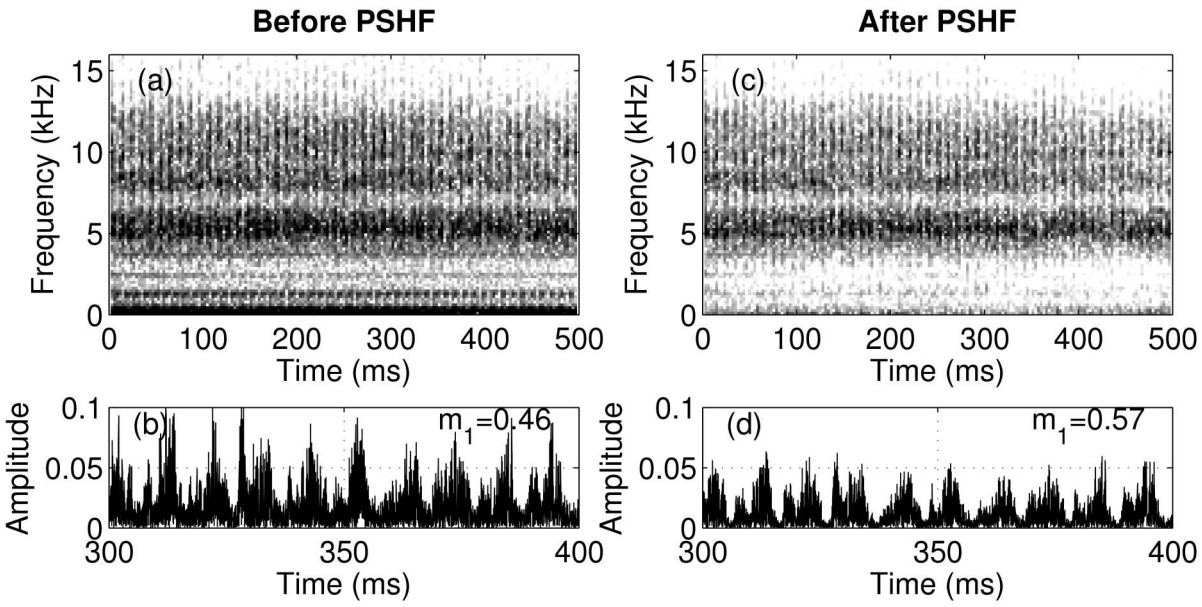

Figure 3.7: 500 ms section of [z]; f_0 ≈125 Hz. Left column: before PSHF. Right column: after PSHF. (a,c) Fixed grayscale spectrograms. (b,d) f_{HP}= 500 Hz filtered magnitude waveforms, $|x(n)|$, for 300–400 ms portion of signal; $\hat{m}$ estimates obtained as in Sec. II.B.

3.3.3 Pitch-scaled harmonic filtering

Separating periodic and aperiodic energy from a mixed-source signal is not a straightforward signal processing task. For speech signals, Yegnanarayana et al. (1998) and Jackson (2000) have proposed algorithms based on comb-filtering of harmonics using adaptive pitch data. By testing the algorithms on synthetic signals, and through informal listening tests, they have shown that speech can be effectively decomposed into periodic and aperiodic streams. In this study, Jackson (2000)'s decomposition algorithm, the *pitch-scaled harmonic filter* (PSHF, described in detail in Jackson and Shadle (2001)), was adopted as preprocessing to the modulation estimation procedure.

Figure 3.7 shows the effect of applying the PSHF to 500 ms of [z] from the sustained fricative corpus. In spectrograms before and after, (a) and (c), the effects of pitch-scaled filtering are evident — formants below 4 kHz have been removed, although there remains some trace of the voicing fundamental.

In Figs. 3.7(b) and (d), the effect on modulation depth of HP filtering employed alone is compared to the combination of PSHF and HP filtering. The HP filtered magnitude waveform, $|x(n)|$, from the PSHF's aperiodic signal (Fig. 3.7(d)) shows deeper and sharper modulation. This was confirmed by measurements of m_1 which gave an increase

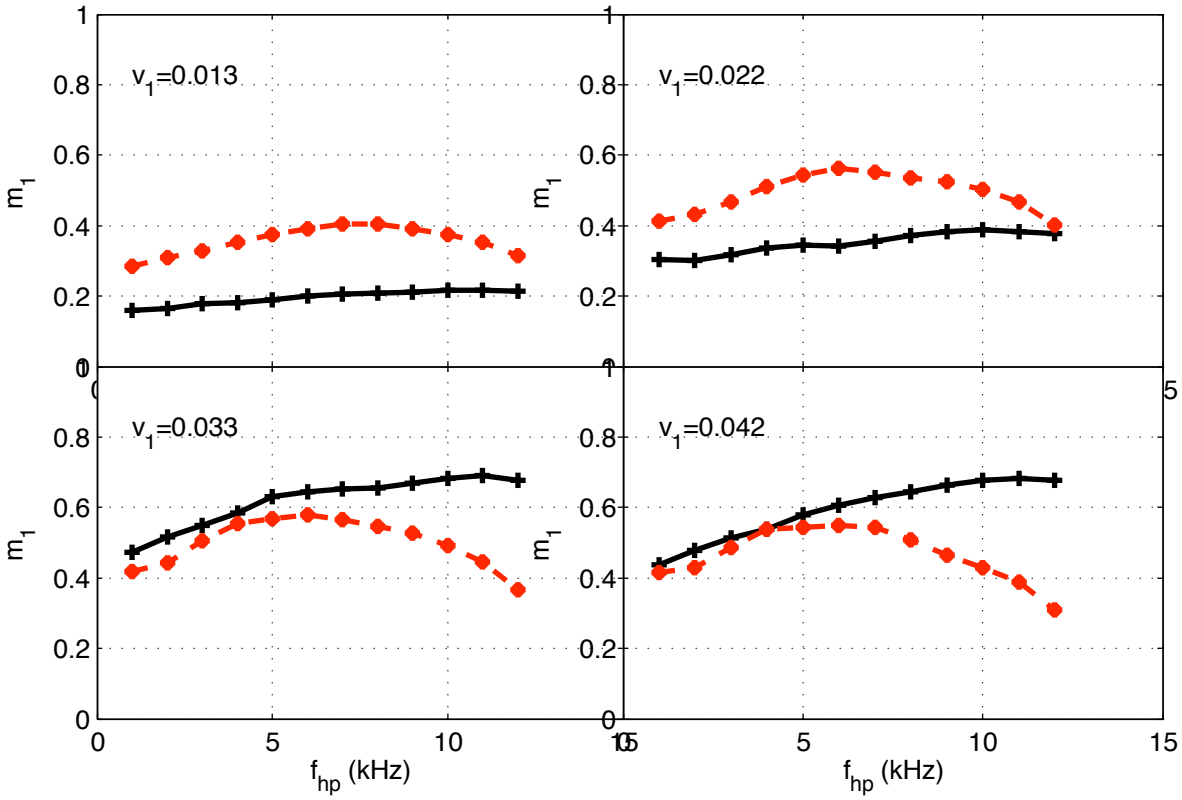

Figure 3.8: Modulation depths, $\hat{m}_1$, as a function of high-pass filter cut-off, f_{HP}, for sustained (solid line) and fluent-speech (dashed line) fricatives. $\hat{m}_1$ values are based on the mean of all readings (across all speakers, fricatives and repetitions) falling within a 0.05 Pa wide bin centred at 0.013 Pa (top left), 0.022 Pa (top right), 0.033 (bottom left) and 0.042 Pa (bottom right).

of 0.11 (from 0.46 to 0.57) after application of the PSHF. The increase is attributed to the attenuating effect of periodic energy on modulation, described in Section 3.3.1.

3.3.4 Processing conditions

HP Cut-off Frequency

To ensure complete removal of the fundamental, high-pass filtering at a low frequency was applied in addition to the PSHF.

The most appropriate f_{HP} represents a balance between excluding as much periodic energy as possible and eliminating too much of the frication to be measured. No single f_{HP} will apply perfectly to every fricative, but averages over the whole data set show the overall effect of filtering. Figure 3.8 shows the effect on modulation depths, when

f_{HP} is varied, at four voicing strengths $v_1 \in \{0.013$ Pa, 0.022 Pa, 0.033 Pa and 0.042 Pa$\}$.

For the two lower voicing strengths in Figure 3.8 (top row), $\hat{m}_1$ for fluent-speech fricatives is slightly above that of the sustained fricatives; the reverse is true of the two higher voicing strengths (bottom row). This reflects the tendency for $\hat{m}_1$ to reach saturation slightly faster in the fluent-speech case, but to settle at a marginally higher level in the sustained fricatives case (covered in the 'Results' section). Of primary interest is the response of $\hat{m}_1$ to changes in f_{HP}: for sustained fricatives, mean $\hat{m}_1$ rises monotonically with f_{HP}; for fluent-speech fricatives, mean $\hat{m}_1$ rises when f_{HP} is in the range 0–5 kHz, but then falls gradually as f_{HP} rises, reflecting the fact that artificial, sustained fricatives have heavily modulated noise in higher spectral areas where little noise is usually found for fricatives produced under conditions of normal articulatory effort.

It is clear the discussion presented in 3.3.2 and from the simulations that it is preferable to use as low an f_{HP} as possible (1 kHz in this case) in order that the AM estimate accurately reflect the modulation of the turbulence noise as a whole.

Window Size

Choice of window size is a trade-off between modulation depth resolution and time resolution, which affects variability such as from pitch glides. Simulations using synthesised signals evaluated different window sizes (see Table 3.1). So, m_1 was estimated with a 100 ms window and a 5 ms step size for the sustained fricative corpus; for the fluent fricatives, a shorter 30 ms window was used. Processing windows were zero-padded to $N = 2^{15}$ points. The required values of f_0 were obtained from analysis of the EGG signal, when available; otherwise, from the speech signal.

Table 3.1: Estimation errors (bias, deviation) over all frames in 100 files versus analysis window size, with 8× zero padding. Values are averaged across modulation index $m \in \{0.0, 0.1, \ldots, 1.0\}$.

f_0	Jitter	Window size	
(Hz)	(%)	1024 (23 ms)	4096 (93 ms)
150	0.0	-0.004, 0.037	0.003, 0.020
160–140	0.5	-0.005, 0.037	0.006, 0.019
180–120	1.5	-0.017, 0.039	0.003, 0.020

3.3.5 Evaluation of modulation estimates

In estimating the underlying modulation depth for a section of voiced frication, errors come from three sources. Error A is due to the nature of the noise signal: random variation inevitably gives a small internal modulation component. Error B is introduced by the modulation estimation procedure (Sec. 3.2.3), as a kind of bias. Finally, in the case of real VFs, imperfections in the preprocessing (Sec. 3.3.3) will introduce further artefacts, error C. Simulation tests were conducted to evaluate the magnitude of the combined estimation error $A + B$. These tests involved making estimates of the modulation index from Gaussian white noise samples with an imposed amplitude modulation.

Summary results for two window sizes are given in Table 3.1 under three voicing conditions, incorporating descending pitch glides and random pitch variation, or jitter. Errors between true and estimated values are given in terms of average bias and variance, quoted as standard deviation. In all cases, the bias was small compared to the deviation, which was twice as high for the short (23 ms) window as for the longer (93 ms) window. The longer window gave errors of ± 0.04 (2σ) on the estimates under typical speech conditions.

Establishing the magnitude of error C is less simple. Filtering partially fills in 'valleys' in the temporal waveform and thus reduces in modulation depth. Eddins (1993) ran simulations to evaluate the effect of band-pass filtering on m_1 of modulated white noise varying the bandwidth, $f_{\mathrm{BW}} \in \{\, 0.2,\, 0.4,\, 0.8,\, 1.6\}$ kHz. He concluded that modulation depth was 'relatively unaffected' for these filter conditions. Our own simulations investigating the effects of limiting bandwidth of modulated noise by high-pass filtering showed the effect to be secondary, increasing the range to ± 0.05 at the highest 11-kHz cut-on frequency (lowest bandwidth). The 1-kHz HP filter applied here has negligible effect, as does the erroneous removal of some noise by the PSHF.

To validate use of the PSHF on VFs, its effect with known modulation was assessed. Phonetically-trained subjects recorded voiced and noise components of VFs separately by producing sustained *voiceless* fricatives, introducing phonation, then gradually relaxing the constriction, leaving just voicing.

Recordings were edited to give voicing plus frication noise with an imposed m. Random, 100 ms sections of frication with known m (0.1–1) were mixed with sections of voicing (from same speaker/fricative) with amplitude varying 0–15 dB in comparison to the frication (periodic to aperiodic ratio, PAR) and pre-processed (Sec. 3.3.3) before measurement of m.

PAR significantly affected the accuracy of estimation for each preprocessing stage. For strongly VFs, the error with HP filtering was much improved by applying the PSHF. Where the voicing component was insignificant, HP filtering produced a better estimate alone, due to PSHF artefacts.

In 1000-trial simulation, where PAR varied freely (as in natural fricatives), overall bias was 0.03, suggesting a tendency to overestimate, and 2σ range rose to 0.10 (cf. 0.18 with HP filtering only). While justifying the use of the PSHF, this result is misleading in some respects. Most VFs are not very strongly voiced, so estimates produced using only HP filtering are fairly reliable; hence accuracy increases only slightly with the PSHF. Tokens with strong voicing, where using the PSHF gave large increases in accuracy, were less common but characteristic of particular speakers or places of articulation. Without the PSHF, results for those speakers and fricatives would be inaccurate, though a fraction of all fricatives. Thus, the PSHF improves comparability of results.

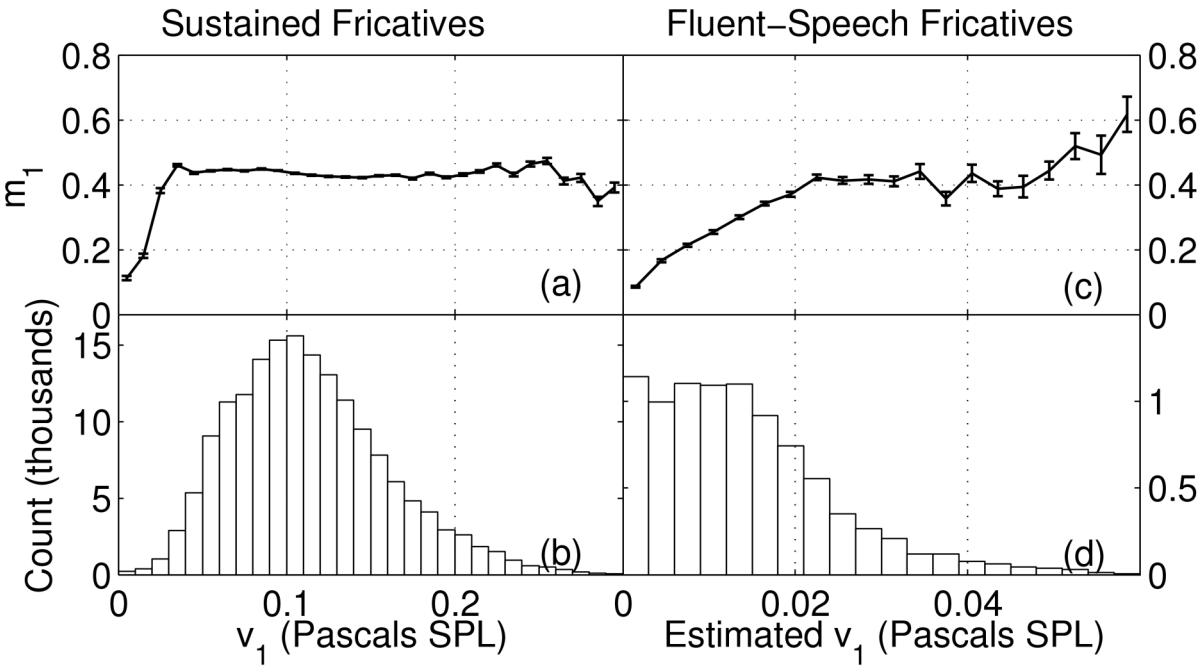

Figure 3.9: Top: Modulation depth $\hat{m}_1$ as a function of voicing strength v_1 or $\hat{v}_1$. Bottom: v_1 or $\hat{v}_1$ distribution histograms for sustained fricatives (left column) and fluent-speech fricatives (right column). Data are means and counts of values falling within ± 0.01 Pa bins (sustained fricatives) or ± 0.003 Pa bins (fluent-speech fricatives). Error bars show standard error.

3.4 Modulation Results

3.4.1 The $\hat{m}_1$ vs. v_1 relationship

Voicing strength v_1 was defined as the spectral amplitude at f_0 in the audio signal prior to high-pass filtering. For sustained fricatives, where subjects' lip–microphone distance was strictly controlled and the microphone calibrated, v_1 is expressed as SPL (in Pa). For fluent-speech fricatives, the calibration to SPL was estimated by comparing RMS measurements averaged over all fluent-speech fricative waveforms to a calibrated test utterance recorded with the sustained fricatives. This estimated voicing strength $\hat{v}_1$ acts as a guide for comparing results from the two experiments.

Figure 3.9 summarises $\hat{m}_1$ for all the data. To explore the relationship between voicing strength, v_1 or $\hat{v}_1$, and modulation depth $\hat{m}_1$, v_1 ranges spanning all the data (0–0.3 Pa SPL for sustained fricatives; 0–0.07 Pa SPL for fluent-speech fricatives) were split into equal bins (0.01 and 0.003 Pa bin width for sustained and fluent-speech respectively). The $\hat{m}_1$ vs. v_1 relationship is represented as voicing-strength bin centres plotted against

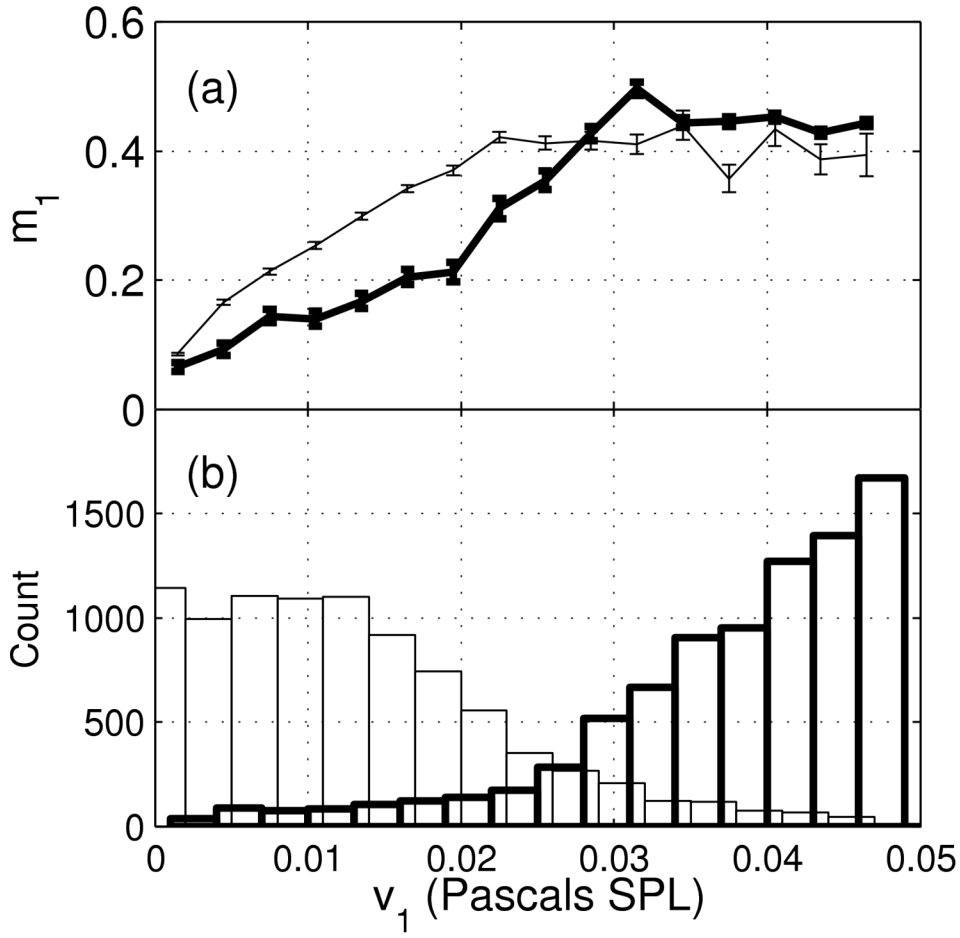

Figure 3.10: (a) Modulation depth $\hat{m}_1$ as a function of voicing strength v_1 or $\hat{v}_1$, and (b) v_1 or $\hat{v}_1$ distribution histogram for sustained fricatives (thick line) and fluent-speech fricatives (thin line). Data are means and counts of values falling within ± 0.003 Pa bins. Error bars show standard error.

average $\hat{m}_1$ reading for that bin. Histograms show number of frames in each bin.

In producing the sustained fricatives, very high or low levels of voicing were seldom used, resulting in an approximately normal distribution. Voicing levels in the fluent-speech case were significantly lower, as expected for short, intervocalic fricatives. The skew of the distribution toward lower values of $\hat{v}_1$ in Fig. 3.9(d) can be attributed to voice dynamics in intervocalic VFs: voicing rapidly decreases in amplitude as frication begins and either remains low until the vowel onset, or ceases (devoicing) (Pincas, 2004).

Figure 3.10 focuses on the low voicing strengths ($0 \leq v_1$ or $\hat{v}_1 \leq 0.05$). There are fewer data frames for the fluent-speech fricatives as each was so short; at higher values of $\hat{v}_1$

where $\hat{m}_1$ was stronger, the lack of data leads to wide error intervals, compared with the sustained fricatives. The $\hat{m}_1$ vs. $\hat{v}_1$ curve for sustained fricatives levels off sharply at $v_1 = 0.03\,\text{Pa}$, where modulation saturates, $\hat{m}_1 \approx 0.5$. Above $v_1 = 0.04\,\text{Pa}$, $\hat{m}_1$ remains constant until $v_1 = 0.25\,\text{Pa}$ (Fig. 3.9(a)), where the data become too sparse to give meaningful results. For fluent-speech fricatives, $\hat{m}_1$ saturated earlier, by $\hat{v}_1 = 0.02\,\text{Pa}$, and was slightly lower ($\sim$0.4) than for sustained fricatives. Above $\hat{v}_1 = 0.03\,\text{Pa}$, data was sparse (Fig. 3.10(b), histogram counts fall below 250) and the bin averages beyond $\hat{v}_1 = 0.05\,\text{Pa}$ should be interpreted with caution.

Figure 3.11 (thick lines) illustrates the $\hat{m}_1$ vs. v_1 relationship for individual speakers. In sustained fricatives, saturation occurred at a similar point (0.03–0.04 Pa) for all subjects except MD; saturation values of $\hat{m}_1$ were also similar for each speaker; quoted $\hat{m}_1$ readings were at 0.055 Pa, from the bin following saturation. Although mean $\hat{m}_1$ ranged 0.13–0.64, the distribution (overall mean = 0.43, std. dev. = 0.12) shows that, on average, speakers' modulation tends to lie around the 0.4–0.5 mark.

Given the imbalance of male to female subjects, only cautious comment can be made in comparison of their results. No difference is immediately discernible in $\hat{m}_1$ at saturation, although statistical comparison reveals a slight difference in mean and distribution (male: mean = 0.40, std. dev. = 0.12; female: mean = 0.50, std. dev. = 0.05).

Individual differences in degree of modulation could correspond to an aspect of voice quality. Significantly, the limiting values of $\hat{m}_1$ came well before modulation was complete ($m_1 = 1$), and imply saturation of a physical AM mechanism.

For the four speakers who took part in both experiments (JP, PJ, AT and RG; two male, two female), comparison of results suggests similar behaviour across experiments (except JP, whose patterns for sustained and fluent-speech fricatives are obviously different). The fluent-speech curves for subjects PJ and RG appear to match the initial portions of their respective sustained fricative curves well. AT's fluent-speech and sustained fricative data complement one another, providing reliable readings at lower voicing strengths and a continuing pattern at higher strengths respectively.

What does the $\hat{m}_1$ vs. v_1 relationship tell us about the distribution of m in typical fricatives? In fluent speech VFs, voicing amplitude tends to be at its highest at the margins of the sound, when frication is weak, and tends to dip at the center of the sound, where frication is strongest. Given the voicing-modulation relationship, m also follows this pattern, with strongest modulation at the onset and offset of the frication noise (Pincas, 2004). Keeping this in mind, for fluent-speech fricatives, peak modulation averaged over all places of articulation clusters around $m \approx 0.5$ (-6 dB). Averaged over the whole fricative, the new modulation depth is more modest, $m \approx 0.35$ (-11 dB).

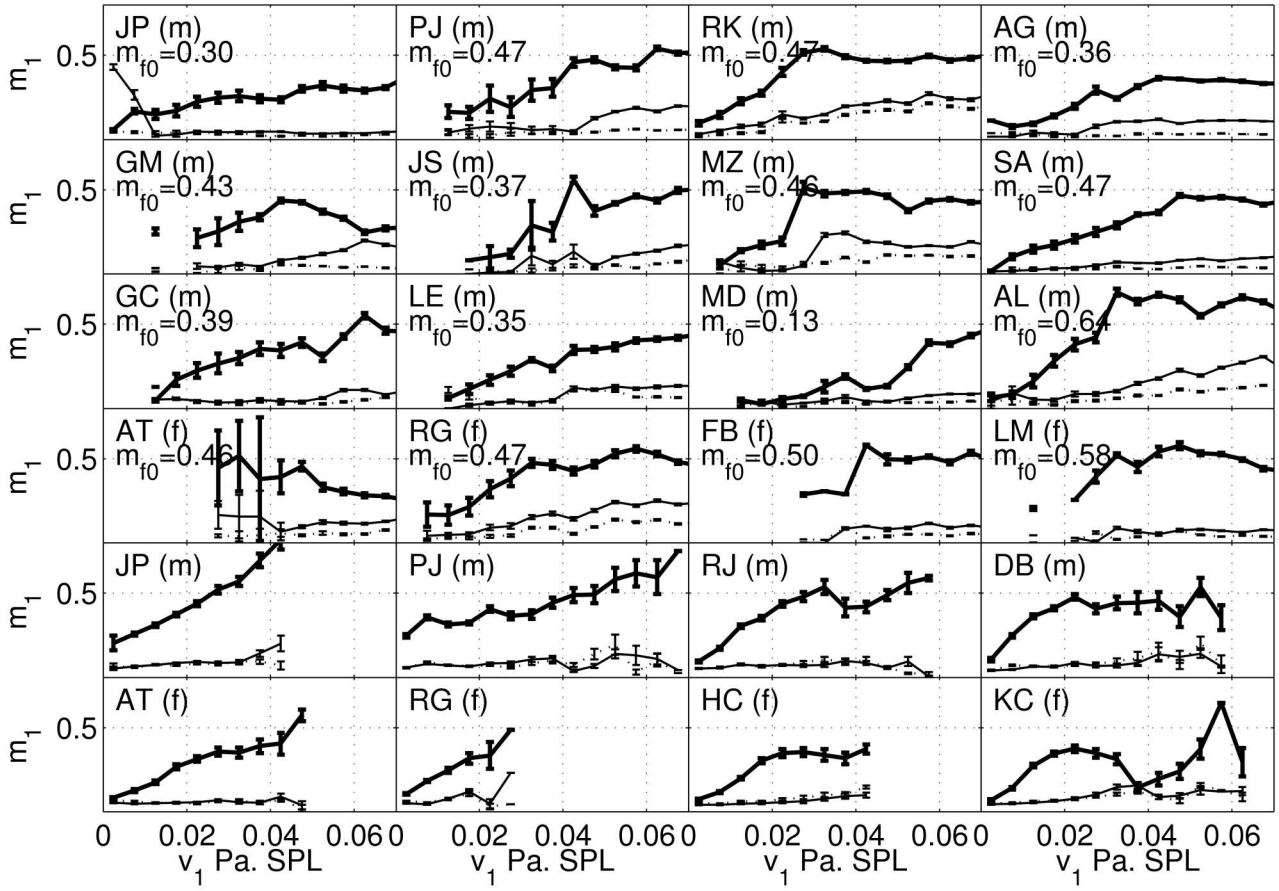

Figure 3.11: Modulation depths at the fundamental frequency $\hat{m}_1$ (thick line), second harmonic $\hat{m}_2$ (thin line) and third harmonic $\hat{m}_3$ (dashed line), versus voicing strength v_1 or $\hat{v}_1$ for individual speakers for sustained fricatives (top four rows) and fluent-speech fricatives (bottom two rows). Data are means and counts of values falling within $\pm 0.005\,\mathrm{Pa}$ bins. Error bars show standard error. Subjects' initials with male/female indication are given. m_1 values quoted for sustained fricatives are mean $\hat{m}_1$ over the voicing strength bin $0.05 \leq v_1 < 0.06\,\mathrm{Pa}$.

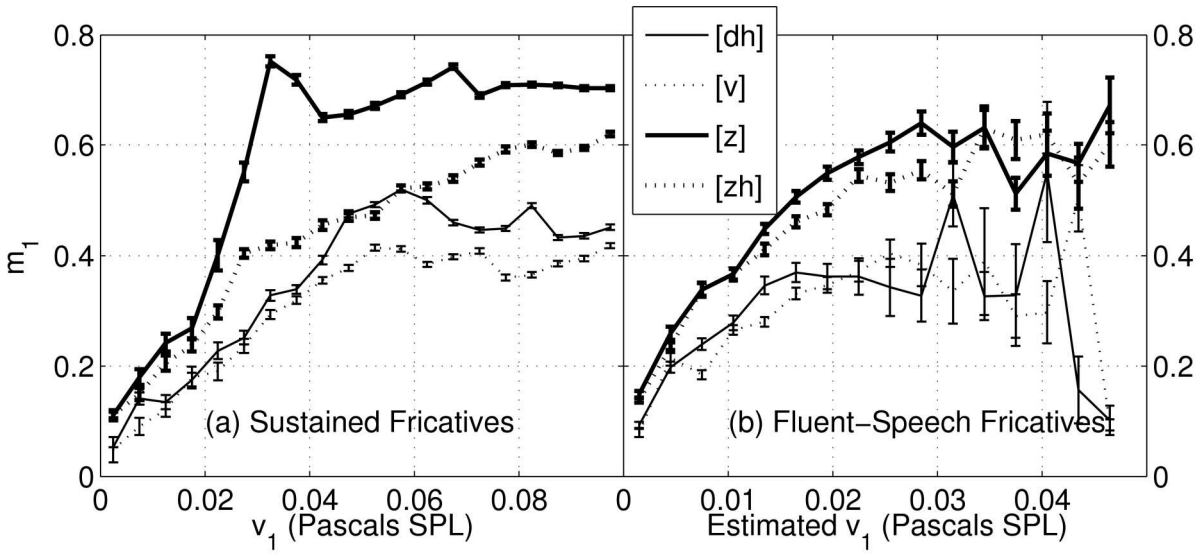

Figure 3.12: Modulation depth $\hat{m}_1$ as a function of voicing strength v_1 or $\hat{v}_1$ for (a) sustained, and (b) fluent-speech fricatives: [ð] – solid thin; [v] – dotted thin; [z] – solid thick and [ʒ] – dotted thick. Data are means and counts of values falling within ± 0.005 Pa bins (sustained fricatives) or ± 0.003 Pa bins (fluent-speech fricatives). Error bars show standard error.

3.4.2 Effect of place of articulation

Modulation depth $\hat{m}_1$ as a function of voicing strength v_1 for fricatives [v] (thin blue dashed); [ð] (thin green solid); [z] (thick red solid) and [ʒ] (thick black dashed). Data are means across speakers, pitches and repetitions. Binning and error bars as per Fig. 3.9.

Differences among the four English VFs are seen in Figure 3.12. Error intervals are wider than those in Figures 3.9 and 3.10 but the basic $\hat{m}_1$ vs. v_1 relationship remains the same for all four fricatives, with varying saturation parameters for each PoA. The curve for [z] (thick solid line) stands out: it is the quickest to saturate (at $v_1 \approx 0.035$) and does so at a highest modulation depth. Furthermore, the transition from the rising, linear part of the curve to the saturated part is more abrupt than for other fricatives. The high modulation depth at saturation for [z] in Fig. 3.12 is common to most speakers: 14 of 16 subjects have [z] as the most heavily modulated fricative at $v_1 = 0.05$ Pa.[4]

[4]In contrast, saturation points and levels for the remaining fricatives, whilst relatively similar and consistently distinct from [z], vary for each speaker with no clear pattern. This could be explained by articulatory configurations varying less across speakers for [z], but more for the other fricatives which tend either to cause difficulty (e.g., [ʒ] is quite rare in English) or to be produced in a variety of ways (e.g., [ð] varies in degree of tongue protrusion). The slightly narrower confidence intervals for [z] at higher voicing strengths concur.

These findings echo previous results for [z] in fluent speech (Pincas and Jackson, 2004). Considering the alternative views of modulated noise production discussed in the Introduction, there are several possible interpretations. According to the static view, the constriction area, A, determines the pressure drop across the constriction, ΔP_C, relative to that at the glottis (Stevens, 1971). So, for [z], which has a marginally smaller constriction (0.17 cm^2) compared to other places (0.19cm^2) (Narayanan et al., 1995), the modulation of ΔP_C, and hence of the flow velocity and noise intensity, would be lesser ($m \sim 0.6$). However, area differences may not be the most significant factor. The monopole, quadrupole and dipole sources for each place of articulation have varied amplitudes and critical Reynolds numbers due to their particular geometry, which could account for the observed differences in m.

The view based on forced turbulence has the advantage that the greater acoustic pressure fluctuation in the smaller constriction would strengthen forcing, tending to raise noise modulation. Yet the precise geometry could have a more substantial influence, for the reasons above, but also since the constriction-obstacle distance and Strouhal number are critical for this mechanism. Modulation is maximal 2–6 diameters from the jet exit, i.e., 1–3 cm, and forcing closer to the natural Strouhal number can double the modulation (Crow and Champagne, 1971). Furthermore, the distribution of sources (e.g., dipoles along the upper lip in non-sibilants [v,dh]) affects modulation phase ϕ_h through turbulence convection (Coker et al., 1996). Thus distributed sources exhibit reduced modulation. Note that alveolar fricatives have the most concentrated dipole source at the lower incisors.

3.4.3 Harmonic structure of $a(n)$

The aeroacoustic processes that produce AM noise in VFs might be thought of as follows: a forcing glottal wave, $d(n)$, interacts with a noise generation process to produce AM noise near the fricative constriction. Following reflections within the VT, the noise radiates as the VF signal, $x(n)$=$a(n)w(n)$. The shape of $x(n)$'s envelope is described by the modulating signal $a(n)$ applied to an unmodulated frication noise signal $w(n)$ and its modulation spectrum has a component m_1 at the fundamental. In relating $d(n)$ to $a(n)$, the results discount the linear hypothesis that $d(n)$ is proportional to $a(n)$ (i.e., that the underlying modulation is identical in shape to the glottal wave that initiated it). This is demonstrated by the saturation of $\hat{m}_1$, the fundamental component of $a(n)$, as a function of v_1, the fundamental component of $d(n)$. Yet, the full $d(n)$ to $a(n)$ mapping requires further clarification.

Observations confirm that even the most strongly modulated frication noise shows negligible components above the second harmonic (i.e., only m_1 and m_2 are significant) and in many cases m_2 is so weak as to blend into the background fluctuations, leaving m_1 only. This is true even when the forcing wave shows significant harmonic structure. Figure 3.4 gives an example of such a situation for a token of [z] taken from the corpus: the forcing wave $d(n)$ is represented by the low-pass filtered audio waveform. This is compared to the high-pass filtered magnitude waveform $|x(n)|$, whose spectrum has peaks at harmonics of the modulating signal $a(n)$. Note how the harmonic structure of $d(n)$ in Fig. 3.4(b) was not preserved in the modulation spectrum of the noise, shown in Fig. 3.4(d).

Figure 3.13 shows $\hat{m}_h$ values at the first and second harmonics using the familiar binning procedure. As v_1 increases, a significant modulation harmonic $\hat{m}_2$ does arise and $\hat{m}_3$ was detectable. Although the results cannot rule out the possibility that m_2 was caused by the same harmonic in the forcing wave (i.e., v_2), it seems more likely that they conform to the behaviour observed by Crow and Champagne in a comparable study using turbulent jets forced by a *pure sinusoid* from a loudspeaker (Crow and Champagne, 1971).

Figure 3.11 shows the harmonic analysis for individual subjects. Some speakers (cf., JP-LM and MZ-RG) show relatively little modulation at the higher harmonics. To ascertain whether this difference depends on the forcing wave's harmonics (voice quality variation), or on natural variation in the modulating signal, requires further investigation.

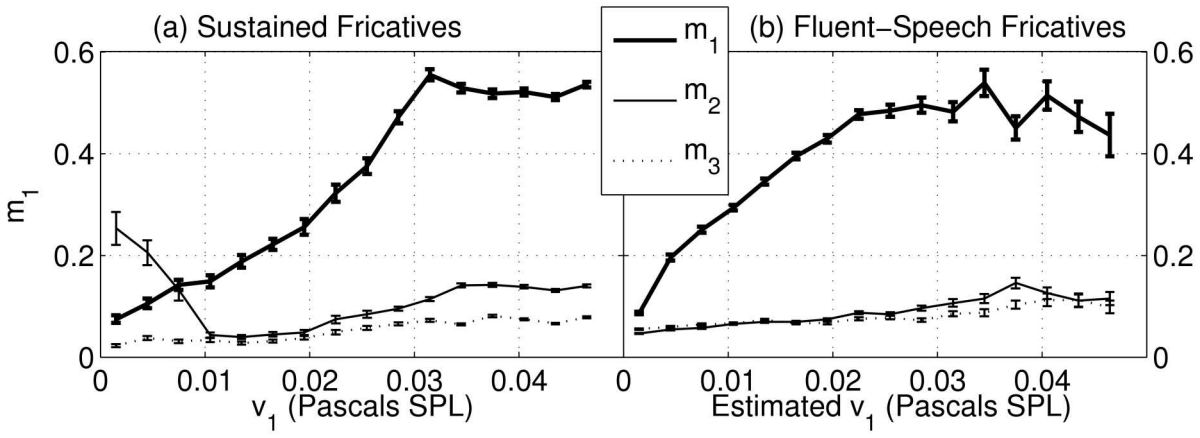

Figure 3.13: Modulation depths at the fundamental frequency $\hat{m}_1$, second harmonic $\hat{m}_2$ and third harmonic $\hat{m}_3$ versus voicing strength v_1 or $\hat{v}_1$ for (a) sustained fricatives, and (b) fluent-speech fricatives. Means from all tokens. Data are means and counts of values falling within ± 0.003 Pa bins. Error bars show standard error.

3.4.4 Effect of f_0

Figure 3.14 analyses the effect of voicing pitch on modulation depth for male and female subjects for both experiments. The relationship between voicing strength, v_1 or $\hat{v}_1$, and modulation depth $\hat{m}_1$ is plotted in Figs. 3.14(a,b,d,e) grouped by fundamental frequency f_0 (bin edges determined by dividing the range of 95% of the data into three equal-width bins). The measured distributions of f_0 are shown in Figs. 3.14(c) and (f).

Figure 3.14(c) reveals that subjects were not very successful in attaining the required f_0 (125, 150, 175 Hz), in the sustained fricatives experiment. Female subjects, as might be expected, had particular difficulty with the lower pitches. The distribution of f_0 data is thus wider than anticipated, but nevertheless provides an appropriate base for analysis. In the fluent-speech fricative experiment, where subjects spoke at their natural pitch, f_0 distributions are significantly tighter. As a result, data are sparse in the lower pitch bins from female subjects (150–180 Hz and 180–210; Fig. 3.14(e)), and dominated by one subject at higher voicing strengths, producing an anomalous curve (KC in Fig. 3.11, bottom right).

Fundamental frequency of voice has little consequence for the relationship between voicing strength and modulation depth, with similar shaped curves throughout. Furthermore, there is some suggestion in the sustained fricative experiment that male subjects (Fig. 3.14(a)) produce higher modulation at lower f_0 for all but one voicing level. However, this pattern is not replicated in any other results and we conclude that f_0 is not an important influence on modulation depth.

3.5 Summary

In VFs, phonation induces AM of frication noise. A technique was developed to estimate the depth of modulation and applied to frication noise from sustained and fluent-speech fricatives. Modulation depth rose approximately linearly with voicing strength for low voicing levels (below ∼63 dB SPL); it saturated at a similar voicing level for different fricatives and speakers, although its value at this point varied. For example, modulation depth at a voicing strength of 0.04 Pa SPL (immediately after saturation) was largest for [z] (0.65; cf. 0.44 for [ʒ], 0.37 for [ð], 0.34 for [v]). Mean m across all fluent-speech fricatives and their corresponding patterns of voicing is ∼0.35, whilst mean peak modulation is higher, at ∼0.5.

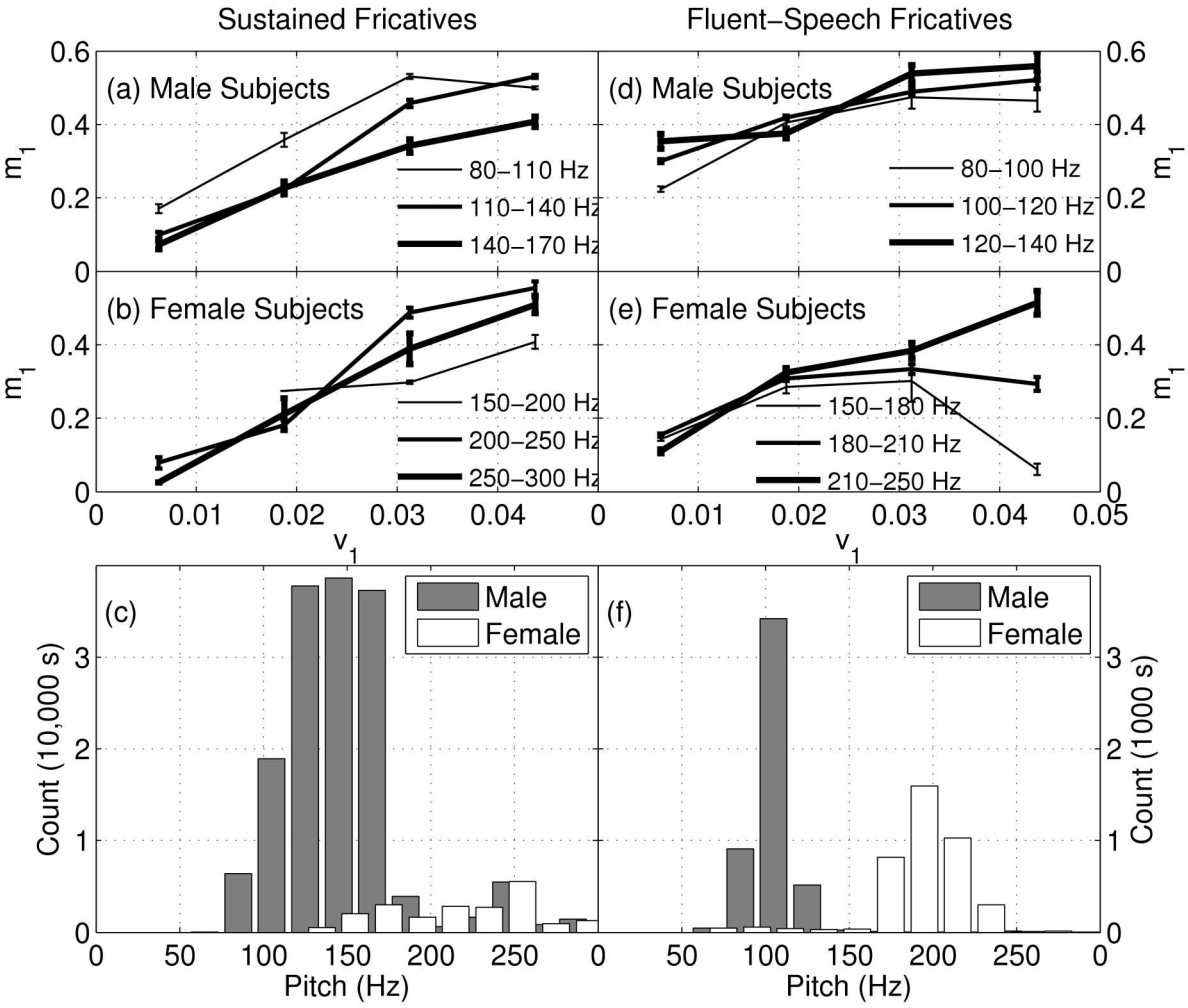

Figure 3.14: Top: Modulation depth $\hat{m}_1$ as a function of voicing strength v_1 or $\hat{v}_1$ for (a) sustained fricatives, male subjects; (b) sustained fricatives, female subjects; (d) fluent-speech fricatives, male subjects; (e) fluent-speech fricatives, female subjects. f_0 data divided into 3 equally-spaced pitch bins (different for each plot). In general: low range (thin line), middle range (medium line), and high range (thick line). For specific bin values see legends. Data for each f_0 bin are means of all frames whose measured f_0 falls into that bin. Voicing strength, v_1 or $\hat{v}_1$, binning used ± 0.005 Pa bins. Error bars show standard error. Bottom: measured f_0 distribution histograms for (c) sustained fricatives, and (f) fluent-speech fricatives. Data are means and counts of values falling within ± 20 Hz bins from all tokens for male (grey bars) and female (clear bars) speakers.

Chapter 4

Psychoacoustic Experiments

4.1 Introduction

AM at the levels measured in the acoustic study ($m \approx 0.35 - 0.5$) would be well above the detection thresholds reported in the psychoacoustics literature and suggest that AM should be perceptible in VFs and available for listeners to use as a speech cue. However, the multitude of complex acoustic features present in VFs precludes a simple interpretation, as suggested in Section 1.1.4.

In this section, five experiments studying AM detection in noise with a simultaneous periodic component are reported. Secondary parameters reflecting specific acoustic characteristics of VFs were also investigated in an approach designed to identify which, if any, of the properties of VFs could inhibit AM detection, leading to its unavailability as a cue to the voicing distinction.

Experiments 1 and 2 study how the relative loudness ('Tone-to-Noise Ratio', henceforth TNR) and phase (ϕ) of a pure tone ($f_0 = f_m$) affects AM detection, in a similar style to the 'distortion tone' experiments of Strickland and Viemeister (1997) and Wiegrebe and Patterson (1999). The initial stimulus conditions were intended to crudely simulate the periodic/aperiodic sound combination in VFs. Experiment 1 has silent gaps between trial intervals, meaning that the underlying tone turns on and off with the noise carrier, reflecting the conditions in isolated fricatives. Experiment 2 presents a continuous tone throughout the 3-stimulus interval, simulating more closely an intervocalic fricative environment and providing a first insight into the importance of context on AM detection.

Both experiments 3 and 4 test the effect on detection of shortening the stimuli from a baseline 500 ms down to 60 ms, a much more realistic VF duration. Experiment 3 also tests the effect on AM detection of shaping the noise with spectral models of typical

fricative sounds taken from speech, giving much more speech-like stimuli. In experiment 4, shorter durations are investigated in combination with the phase relationship between tone and noise modulation to establish whether any effect for phase is also observable at very short durations of noise, as might be found in fluent speech.

Finally, experiment 5 combines a short (80 ms) frication duration with natural voicing (rather than a pure tone) and investigates the effect of context using recorded, naturally spoken vowel environments. Vowel type, relative amplitude ('Vowel-to-Frication Ratio', henceforth VFR) and vowel filtering are the context parameters varied.

4.2 Method

Methods for generating the stimuli differ in some areas between experiments 1–4, which used wholly synthetic stimuli, and experiment 5, which used a combination of recorded audio and synthetic noise.

4.2.1 Stimuli for Experiments 1–4

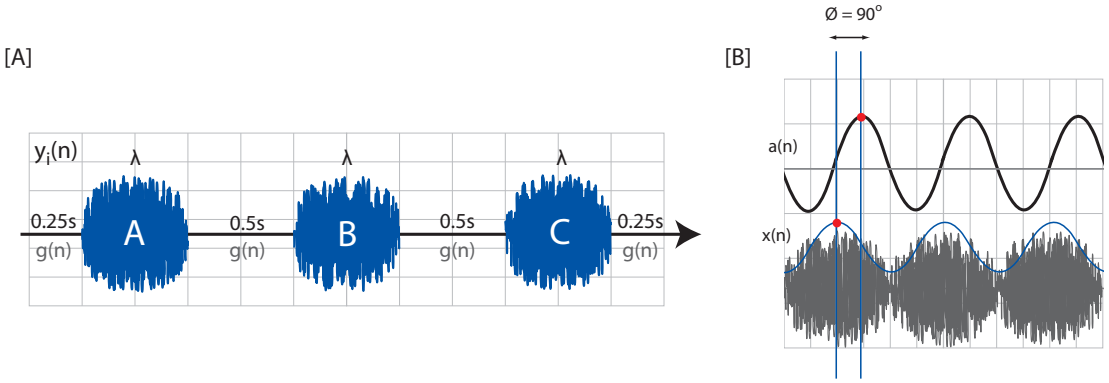

Figure 4.1: A) Schematic illustration of synthesised 3-interval trial, $y_i(n)$, showing interval durations, λ, and inter-interval timings. B) Close-up schematic illustration of ϕ (tone-modulation phase relationship) for 90° condition.

Trials consisting of SAM noise plus sinusoid were pre-generated at $f_s = 32$ kHz with 16-bit resolution and stored for presentation in uncompressed WAV format. As illustrated in Figure 4.1(a), a trial, $y_i(n)$, was composed of 3 noise intervals of duration λ, interspersed with gaps, $g(n)$. Duration of preceding and trailing $g(n)$ was 250 ms and inter-interval $g(n)$ was 500 ms, therefore total trial duration was $3\lambda + 1.5$ seconds. The nature of $g(n)$ depended on experimental condition: in experiments 2, 3 and 4, the signal in $g(n)$ was filled by the tone; in experiment 1, the gaps were silent. In what follows we make reference to a baseline interval which has $\lambda = 500$ ms.

Tone

The tone was a sinusoid of the same frequency, f_0, as that of the modulating signal, f_m. The amplitude of the sinusoid, d, was calculated as $d = 10^{R/20}\sqrt{2}$, where R is the decibel ratio of its RMS amplitude to that of the noise component (which is always 1). R is henceforth referred to as the 'tone-to-noise ratio' or TNR.

For the sinusoid, $a(n)$, a continuous function with duration $3\lambda + 1.5$ was generated as $a(n) = d \sin\left(\frac{2\pi f_0 n}{f_S}\right)$[1]. To produce a smooth rise and decay in the amplitude, the onset and offset (initial 10% rise and final 10% fall) of $x(n)$ were gated by a raised cosine ramp, $h(n)$, e.g., initially for $n = 1..N$:

$$h(n) = \frac{1}{2}\left(1 - \cos\frac{\pi n}{N}\right), \tag{4.1}$$

where, for example, N=800 is the ramp duration in samples (50 ms) for the λ=500 ms baseline interval.

Noise

The required number of samples of noise carrier, $w(n)$, were generated using the *RANDN* command in MATLAB (random entries, chosen from a normal distribution with mean zero and standard deviation one). Pseudo random noise tokens with unusually large crest factor were removed to avoid their risk of identification and bias to the detection results. For experiment 3 noise was shaped (see Figure 4.4), otherwise white noise was used. The noise, $w(n)$, was then modulated as follows:

$$x(n) = w(n)\left[1 + m \sin\left(\frac{2\pi f_m n}{f_S} + \phi\right)\right]\sqrt{1 + m^2/2}, \tag{4.2}$$

where m is modulation index in the range 0 to 1, f_s is the sampling frequency, ϕ is a phase offset, f_m is the modulation frequency (always 125 Hz — around the pitch of a typical male voice), and the final term provides power normalisation[2]. Stimuli were generated for values of m from 0 dB to -25 dB in 0.5 dB steps, and gated by the raised cosine ramp as in Eq. 5.2.6.

[1] At this point, an important methodological point arises. We chose to avoid the traditional method of separately generating the tone and noise sources and combining them at runtime due to possible inaccuracies in the rendition of the phase difference between noise modulation and tone that this might introduce. By generating the stimuli with sinusoid and noise already combined, we were able to guarantee that the required phase difference was accurately reproduced at runtime. Furthermore, the constant-tone experiment required a continuous tone throughout all three intervals of the 3AFC paradigm used (see Section 4.2.2). Preliminary tests showed that this could not be adequately simulated by concatenating individual tone segments, facilitating a similar design for the tone-gap and constant-tone experiments (where identical stimuli were separated by silence in experiment 1 and segments of tone in experiment 2). Instead, for each modulation depth, three entire 3-alternative trials were generated, with the target (modulated) stimulus in all three possible positions. At runtime, the controlling computer chose randomly from these three trials.

[2] The imposition of modulation on $w(n)$ causes an increase in the RMS amplitude of the noise in proportion to $\sqrt{1 + m^2/2}$. To prevent subjects detecting modulation on the basis of increased energy of the target over the standards, stimuli are normalised and the average intensity of the noise component of stimuli with different modulation depths (including the unmodulated standards) was identical.

Combination into trials

For each experiment, three permutations, $y_1(n), y_2(n)$ and $y_3(n)$, were produced; the subscript corresponds to the interval of the modulated target. Noise was always aligned to start from an up-going zero-crossing of the sinusoid. The required phase difference between the modulation and tone was created by the phase variable, ϕ, in Equation 4.2 and thus the modulation began with differing phase leading the tone. This is illustrated in Figure 4.1(b) for the $\phi = 90°$ phase lead condition.

Note that pilot data from Viemeister (1979) indicated that 'the phase of modulation at offset had no measurable effect on modulation threshold'; we assume that phase of modulation at onset also has negligible effect once the initial and final raised cosine ramps have been applied.

4.2.2 Experimental Procedure

A three-alternative forced-choice (3AFC) paradigm with 2-down 1-up procedure and decreasing step-size was adopted. In this way, an estimate of the modulation depth necessary for 70.7% correct responses is acheived (Levitt, 1970). The subject's task was to identify a modulated target interval from the two unmodulated intervals. The order of presentation of the three intervals was randomly chosen for each trial and the start of each interval was cued visually. An on-screen display then prompted the subject to respond and gave feedback as to whether the response was right or wrong.

For each staircase, (e.g., Phase, TNR, VFR), threshold estimation proceeded using a set of staircases as follows. The modulation depth of the target was set to its initial level, determined by informal listening tests. After two subsequent correct responses, it was decreased by one step; after a single incorrect response, it was increased by one step. Step-size started at 2 dB and decreased to 1 dB after 2 reversals, and then to 0.5 dB after a further 2 reversals; 6 reversals were then obtained at this final step size before terminating. Threshold was the average of these final 6 reversals. If the standard deviation of the six final reversals exceeded 3 dB, the threshold measurement was discarded. The number of threshold measurements collected for each condition and subject was between 1 and 5 depending on results and particular experiment.

Adaptation to modulation has been a source of considerable methodological difficulty in modulation *discrimination* experiments in the past. Both Ewert and Dau (2004) and Wakefield and Viemeister (1990), investigating AM discrimination, found that thresholds were substantially raised when the depth of the standard of the staircase was more than 5 dB lower than that of the previous staircase. As Wakefield and Viemeister (1990) comment, "prior exposure to envelopes with large modulation depths

may lead to poorer discrimination of envelopes with small modulation depths". In their work, this led to a modification of randomisation technique, where, instead of completely randomising the order of presentation of staircases with differing standard depths, randomisation took place within groups of staircases where the standard depth differed by no more than 5 dB.

The above problem is not encountered in the current experiment given that we are interested in *detection*, and as such, the standard is always unmodulated noise. However, it does raise the issue of presentation order of staircases. To avoid sequential effects as far as possible, we interlaced staircases for different test conditions. In experiments 1 and 2, the six TNR staircases were run simultaneously and interlaced. In experiments 3 and 4, five duration staircases were interlaced: for example, experiment 1 started with a trial at 20 dB TNR, the subject responded, the position in the adaptive procedure was recorded, another TNR was chosen at random, the position in the adaptive procedure was retrieved and the corresponding trial was presented. This continues until the requisite number of reversals at each TNR had been obtained, yielding one threshold measurement for each TNR staircase.

4.2.3 Presentation of Stimuli

Sound files were played through a PC sound card (Creative Labs Extigy). Mono stimuli were presented to both ears through open-back headphones (Sennheiser HD414) to subjects seated in a double-skinned sound proof chamber. Visual displays were on a computer monitor and subjects' responses were by keyboard. Stimulus presentation, subject response and feedback were controlled by the NBS Presentation 9.30 computer program.

4.2.4 Calibration and Testing of Equipment

A B&K Type 2610 measuring amplifier (22.4 Hz high-pass filtering, 'fast' RMS averaging) was used to calibrate the SPL of the noise portion of all stimuli to 50 dB SPL. Before commencing the experiments, the accuracy of stimuli playback was measured by recording output through the sound card using a B&K instrument microphone. Frequency response of the headphones had some effect on the quality of the noise reproduced, but TNRs were maintained to within 5 dB in all cases. After experiment 1, when the effect of phase was investigated, the temporal accuracy of the PC sound card was checked by displaying the output signal with a Tekno 1200 oscilloscope. It was found that the polarity of the signal was always reversed, causing a 180° offset in reproduced phase; results have thus been adjusted to compensate for this.

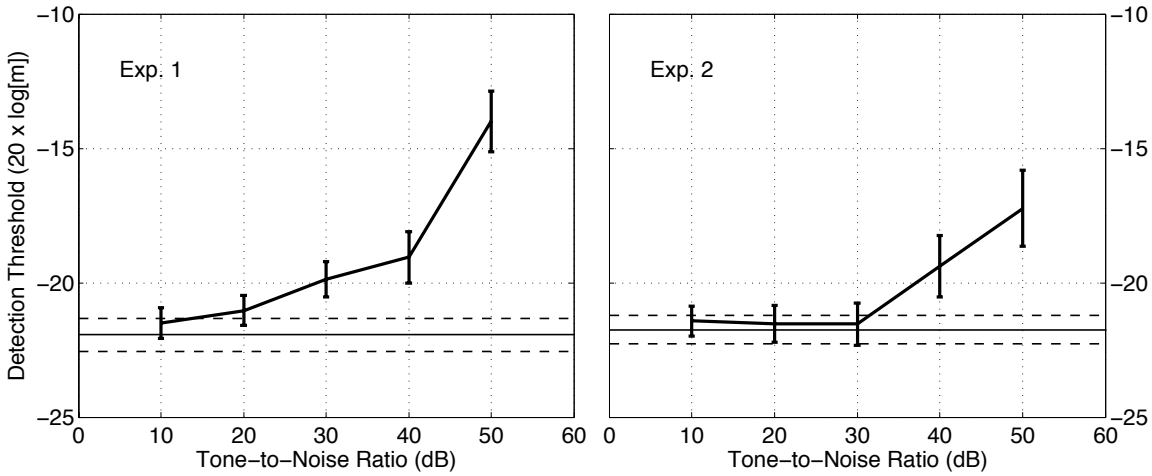

Figure 4.2: Experiments 1 (left) and 2 (right). AM detection thresholds, m_d, as a function of tone-to-noise ratio (TNR). Error bars are 95% confidence intervals. Threshold estimates averaged over all subjects and phase conditions. Solid horizontal line is 'no-tone' condition (dashed horizontal lines show corresponding 95% CI).

4.3 Experiment 1: Tone-to-Noise Ratio

Stimulus Interval Duration	500 ms	
Inter-Interval Gap	500 ms silence	
f_0 of Tone and Modulator	125 Hz	
Tone-to-Noise Ratio (TNR)	10, 20, 30, 40 and 50 dB	(Staircase variable)
Phase (ϕ)	0°, 90°, 180°, 270°	(Scenario variable)
Noise Shape	White noise	
Subjects	3 M, 1 F (Total 4)	
Threshold Runs per Estimate	Min. 3	

Table 4.1: Summary of conditions and parameters for psychoacoustic experiment 1.

AM noise and tone stimuli intervals presented in 3AFC trials were separated by silent gaps. Four subjects were each tested on 6 TNRs at each of four phase settings [3]. For each unique combination of variables, at least 3 threshold measurements were obtained as described in Section 4.2-4.2.2.

A pattern of rising threshold (worse detection performance) as TNR increased was observed for all speakers; this relationship is illustrated in Figure 4.2 (left) as an average across subjects and phase. The performance without any tone ($-\infty$ dB TNR) is shown

[3] Accuracy of phase reproduction was not verified in this experiment, so only averages over phase are reported.

as a solid horizontal line with 95% confidence intervals (CI) as dashed lines either side; threshold estimates for each TNR are displayed with 95% CIs. All threshold estimates were based on at least 48 individual threshold measurements (4 subjects, 4 phases, 3^+ threshold runs).

Detection threshold rises gradually from approximately -22 dB in the No-Tone case to -14 dB for the 50 dB TNR case. A 2-way ANOVA (Subject, TNR) confirms an effect for TNR as expected ($p < 0.00$). Post-hoc tests at the $p < 0.05$ level with Bonferroni adjustment show that the mean at each TNR is not always significantly different from adjacent TNRs: for quiet tones (10–20 dB above noise level) there is no significant increase over base performance but their thresholds are significantly different from mid-volume tones (30–40 dB TNR). The loudest tone (50 dB) gave a significant threshold increase of approximately 5 dB from the 30–40 dB tones (8 dB above base performance).

Although subjects displayed the same pattern of rising thresholds as TNR increased, 2-way ANOVA ($p < 0.05$) reveals a main effect for Subject and a Subject/TNR interaction effect, manifested in slightly higher thresholds for quiet and mid-volume tones for subjects 1 and 4 (non-significant at $p < 0.05$, Bonferroni adjusted) and differences in the size of threshold increase for 40-50 dB TNR tones between subjects: for example, 50 dB TNR thresholds for subjects 3 and 4 differed by 4 dB (significant at $p < 0.05$, Bonferroni adjusted). This difference is apparent in the wider CI intervals in Figure 4.2 (left) at 40-50 dB TNR and suggests that the effect of the louder tone depends somewhat on the subject.

The results of Wiegrebe and Patterson (1999) are hard to compare to our findings as a fixed modulation depth of $m = 0.5$ (-6 dB) and a different experimental technique were used, but in some circumstances a rise of only 6 dB in the level of the tone was enough to move from 100% correct identification of the modulated interval to chance. Wakefield and Viemeister (1985) carried out experiments with fixed noise and variable tone level, and vice versa. In parallel with our findings, a combinations of tone and noise levels equivalent to 25 dB TNR were insufficient to produce their observed phase effect at higher TNRs ($\geq$35 dB TNR). In the next experiment, the effect of phase was investigated.

4.4 Experiment 2: Phase

Stimulus Interval Duration	500 ms	
Inter-Interval Gap	500 ms tone	
f_0 of Tone and Modulator	125 Hz	
Tone-to-Noise Ratio (TNR)	10, 20, 30, 40 and 50 dB	(Staircase variable)
Phase (ϕ)	0°, 90°, 180°, 270°	(Scenario variable)
Noise Shape	White noise	
Subjects	2 M, 2 F (Total 4)	
Threshold Runs per Estimate	Min. 3	

Table 4.2: Summary of conditions and parameters for psychoacoustic experiment 2.

Noise and tone stimuli presented in 3AFC trials were separated by a pure tone in the gaps between intervals. This was intended to maximise exposure to the tone to facilitate subjects' use of phase information and to crudely simulate an intervocalic environment.

A 3-way ANOVA (Subject, TNR and Phase) confirmed main effects ($p < 0.05$) for all factors and Subject/Phase and Subject/TNR interaction effects. Figure 4.3 (individual phase conditions, averaged over all subjects) shows a pattern consistent with experiment 1 for all phases except 0°. The TNR effect was largest for the 180° (out-of-phase) condition: there was significant deterioration from no-tone performance at all TNRs, with the magnitude of the effect increasing from approximately 1 dB at 10 dB TNR to approximately 10 dB at 50 dB TNR. The effect was thus more marked than in experiment 1, where results were averaged across phase. The reverse of this effect is seen for the 0° (in-phase) condition, in which AM detection was significantly *improved* over no-tone performance by the presence of the tone for TNRs in the range 20–30 dB. At larger TNRs, performance returned to approximately the same as the no-tone case.

This difference between the 0° and 180° cases is significant at all TNRs ($p < 0.05$, Bonferroni adjusted for multiple comparisons), whereas the 90° and 270° phase conditions exhibited similar trends, with no significant difference at any TNR. Detection thresholds for these phases lie in between the extreme cases of 0° and 180°, but wider CIs mean the difference between those phases and 90°/270° is only significant beyond 30 dB TNR. AM detection was thus dependent on phase. In most cases, detection was hindered by the presence of the tone (greatest effect observed for out-of-phase tone and modulation), but for in-phase tone and modulation, an *improvement* in detection was achieved.

Wakefield and Viemeister (1985) observed a 180° separation between maximum enhancement and impairment of detection, and the magnitude of the effect increased as

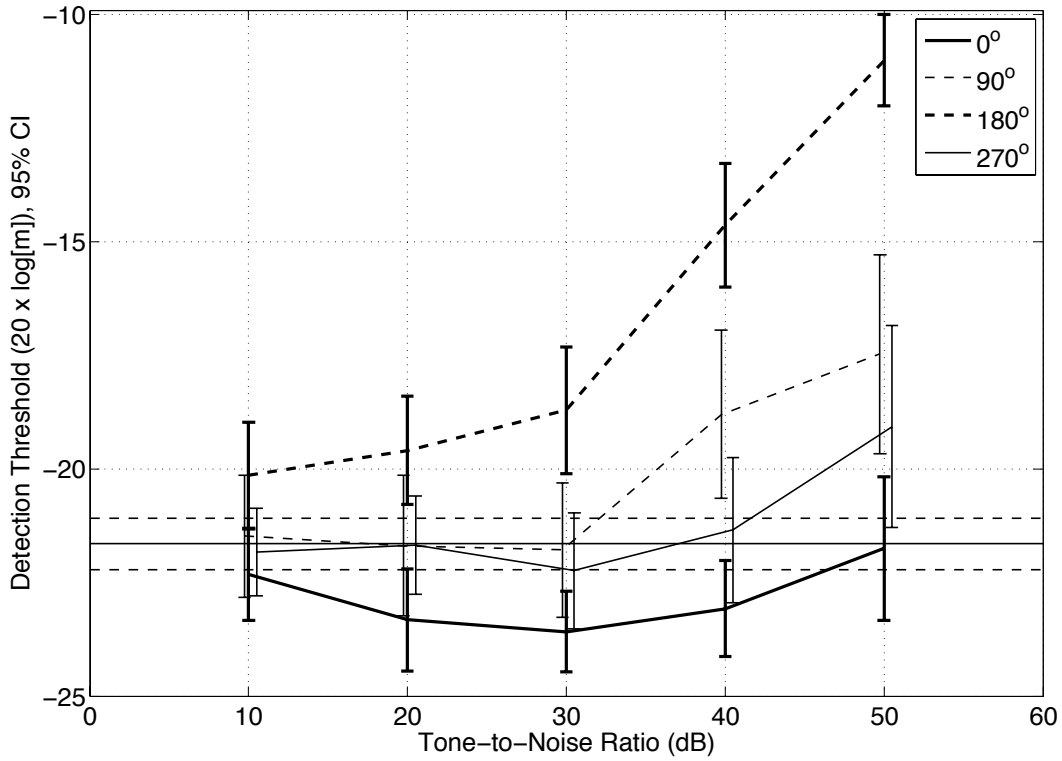

Figure 4.3: Experiment 2. AM detection thresholds, m_d, as a function of tone-to-noise ratio (TNR) for different phase conditions. Error bars are 95% confidence intervals. Results averaged over all subjects. Solid horizontal line is 'no-tone'condition (dashed horizontal lines show corresponding 95% CI).

the level of the tone increased, and vice versa. However, they report optimum detection at approximately 90° and maximum impairment at 270° (in comparison to 0°/180° in this study), but this could be attributable to difference in stimulus generation technique where exact values of ϕ depend on precise combination of noise with the tone.

In keeping with our findings they found that the enhancement/impairment effect was not necessarily symmetrical; they note that "over a limited range of relative phases, the low-frequency tone appears to enhance the sensitivity to amplitude modulation whereas for most other phases, it appears to impair sensitivity". In our results, only the 0° case was able to produce detection enhancement, whereas all other phases impaired detection. Our results corroborate their finding that the magnitude of the detection-enhancement effect at those phases where it occurs (1–2 dB in this study) is significantly smaller than than the impairment effect (5–10 dB in this study). They showed a maximum detection enhancement of approximately 5 dB, but at other phases impairment was so strong for some subjects that it was not possible to measure a threshold. The difference in the nature of the stimuli used in the experiments (broadband carriers in the current work compared to 3 kHz narrowband noise in their case) may explain the difference in the magnitude of the effect. If the responsible mechanism is 'additive' or 'multiplicative' interaction in the region of carrier by the low-frequency tone, the broadband noise used in this study would presumably be much less effectively 'masked' than their narrowband noise, although a significant amount of further study is needed to establish the relationship between tone/modulation frequency and the characteristics of the noise carrier (i.e., bandwidth and center frequency).

Note that our experiment 1 results, and those of Wakefield and Viemeister (1985), tend to impairment of detection over enhancement (both in the range of phases over which the effect is operative and the magnitude of the effect). When averaged over phase, the stronger impairment effect dominates, giving rise to a general pattern of increasing thresholds at higher TNRs (Figure 4.2, right), as was seen in experiment 1 (Figure 4.2, left). Thresholds at 30 dB, 40 dB and 50 dB were higher in experiment 1 than experiment 2, suggesting that enhancement from the 0° phase condition was weaker or absent in experiment 1, possibly due to the gaps in the tone giving listeners less time to attune to phase information. The slightly wider variance in experiment 2 at 40 dB and 50 dB TNR supports the idea of a larger range of threshold measurements caused by a stronger phase effect.

4.5 Experiment 3: Spectral Shape and Duration

Stimulus Interval Duration	500, 160, 120, 80, 60 ms	(Staircase variable)
Inter-Interval Gap	500 ms tone	
f_0 of Tone and Modulator	125 Hz	
Tone-to-Noise Ratio (TNR)	20 dB	
Phase (ϕ)	48°	
Noise Shape	Filtered White noise: [f, θ, s, ʃ]	(Scenario variable)
Subjects	6 M, 6 F (Total 12)	
Threshold Runs per Estimate	1	

Table 4.3: Summary of conditions and parameters for psychoacoustic experiment 3.

Using experiment 2 (constant-tone) as a template (20 dB TNR and ϕ= 48°, values that correspond approximately to VFs in speech), durations of the stimuli and the spectral shape of the modulated noise were varied to investigate their effect on AM detection. The shapes of the noise spectra were based on the four places of articulation for English fricatives. LPC coefficients were estimated using voiceless tokens. Stimulus durations of 60 ms, 80 ms, 120 ms and 160 ms were tested in addition to the 500 ms standard. Broadband noise (as used in experiments 1 and 2) was shaped with simple spectral models of these fricatives constructed by LPC analysis (6^{th} order) of the central portions of fricatives recorded in nonsense words spoken by one of the authors (JP, 16 repetitions of each fricative, manual labelling of central portions). Figure 4.4 illustrates the resulting LPC spectra used to shape the white noise.

Twelve subjects each completed a single threshold measurement for all 4 noise shapes with all 5 durations for each condition. Like the procedure adopted in experiments 1 and 2, each noise Shape was tested separately, interleaving Duration staircases.

Overall, Shape appears to have little effect on either overall thresholds or patterning with Duration: noise-shape means are all within 1 dB of each other and 2-way ANOVA found no Noise/Duration interaction ($p < 0.05$). Results for all noise conditions for further analysis of the duration effect are henceforth combined.

To attain a higher degree of generalisation from the results of this experiment, we increased the number of subjects threefold whilst reducing the number of threshold runs per subject by the same factor. An equivalent amount of data was thus collected in this experiment as in the previous two, although we do not consider subjects' results on an individual basis. Instead, a subject's results for shorter durations (60 ms, 80 ms, 120 ms and 160 ms) were normalised to the average threshold (across noise shapes) that

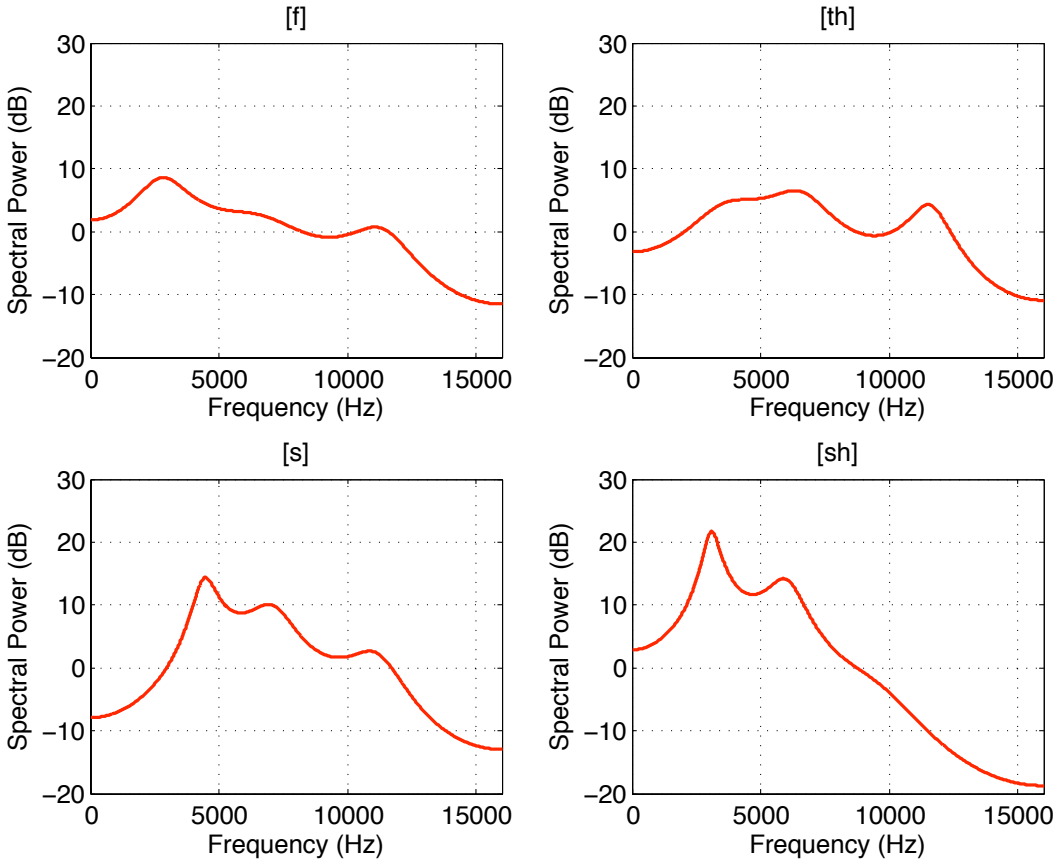

Figure 4.4: LPC spectra used to shape the noise for stimulus generation in experiment 3.

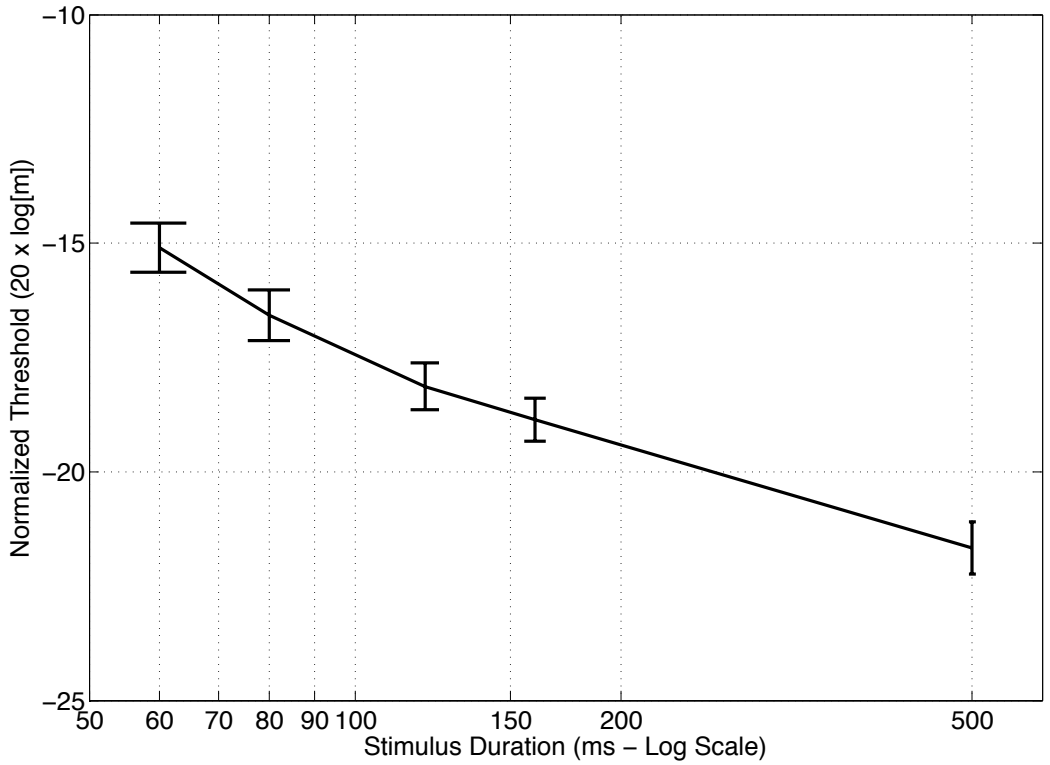

Figure 4.5: Subject normalised AM detection thresholds as a function of stimulus duration averaged over subjects and noise shape. See text for details of normalisation procedure. Error bars are 95% confidence intervals.

he/she obtained for the baseline stimuli (500 ms duration) and a *threshold increase*, Δm_d, was computed. All subjects' Δm_d were then averaged for analysis (i.e., Subject variable was discarded), and offset by the overall average threshold across subjects for the baseline (i.e., mean Δm_d+ -21.9 dB) to give a *normalised threshold*. Results are plotted in Figure 4.5.

Main effect for Duration was significant to the $p < 0.01$ level (2-way ANOVA). Figure 4.5 illustrates the increase in AM detection threshold from the standard (500 ms) case when shortening the stimulus duration. Results for 60 ms, 80 ms and 120 ms are all significantly different from each other ($p < 0.05$, Bonferroni post-hoc tests) and the relationship in the interval is log-log with an approximate 3 dB decrease in AM detection threshold for a doubling of duration (between 60 ms and 120 ms). Between 120 ms and 500 ms, a similar decrease in detection threshold was observed (although the duration was 4 times, rather than twice). Further tests would be needed to establish whether improvements in detection can be achieved for stimuli longer than 500 ms, or indeed to determine the minimum duration required for detection below 60 ms. The overall 7 dB difference that was measured concurs with Viemeister (1979) and Sheft

and Yost (1990); the latter reported a 9 dB change in detection threshold between 25 ms and 400 ms stimuli at m_f=160 Hz.

4.6 Experiment 4: Duration and Phase

Stimulus Interval Duration	500, 160, 120, 80, 60 ms	(Staircase variable)
Inter-Interval Gap	500 ms tone	
f_0 of Tone and Modulator	125 Hz	
Tone-to-Noise Ratio (TNR)	40 dB	
Phase (ϕ)	0° and 180°	(Scenario variable)
Noise Shape	Filtered white noise: [s]	
Subjects	6 M, 6 F (Total 12)	
Threshold Runs per Estimate	1	

Table 4.4: Summary of conditions and parameters for psychoacoustic experiment 4.

To investigate whether there is an interaction between duration and phase, the same set of durations (60 ms, 80 ms, 120 ms, 160 ms and 500 ms) was investigated for a fixed tone level known to produce a strong phase effect (40 dB TNR; the loudest tone was avoided to minimise the possible effect of masking). Two phase settings (0° and 180°) were tested using the [s]-noise spectral shape from experiment 3. These modulation phases were shown to produce the largest spread of detection in experiment 2, and are thus apt for establishing how phase patterns with duration.

Twelve subjects each completed a single threshold measurement for both phases with all 5 durations. Duration staircases were interleaved whilst the two Phase conditions were tested in separate sessions.

A subject normalisation procedure similar to that used in experiment 3 was adopted, using subjects' average threshold (over 0° and 180° conditions) for the 500 ms stimuli to calculate Δm_d which was then averaged over subject and offset by the overall speaker average for baseline (500 ms) detection.

The pattern of rising threshold with decreasing duration was extremely similar for both phases (note the parallel sloping lines in Figure 4.6) and reflected the pattern found in experiment 3. Thus, the magnitude of the advantage to detection offered by the 0° phase over the 180° was maintained for shorter stimulus durations down, to as little as 60 ms. A 2-way ANOVA indicated a lack of main effect and interaction effect for phase when the offset between the two curves was discarded: even for 60 ms, there was only a 0.1 dB difference between normalised threshold increases for 0° and 180° phase conditions.

One might expect similar absolute thresholds for both phases at the shortest durations, where little phase information is available for enhancement or impairment of detection.

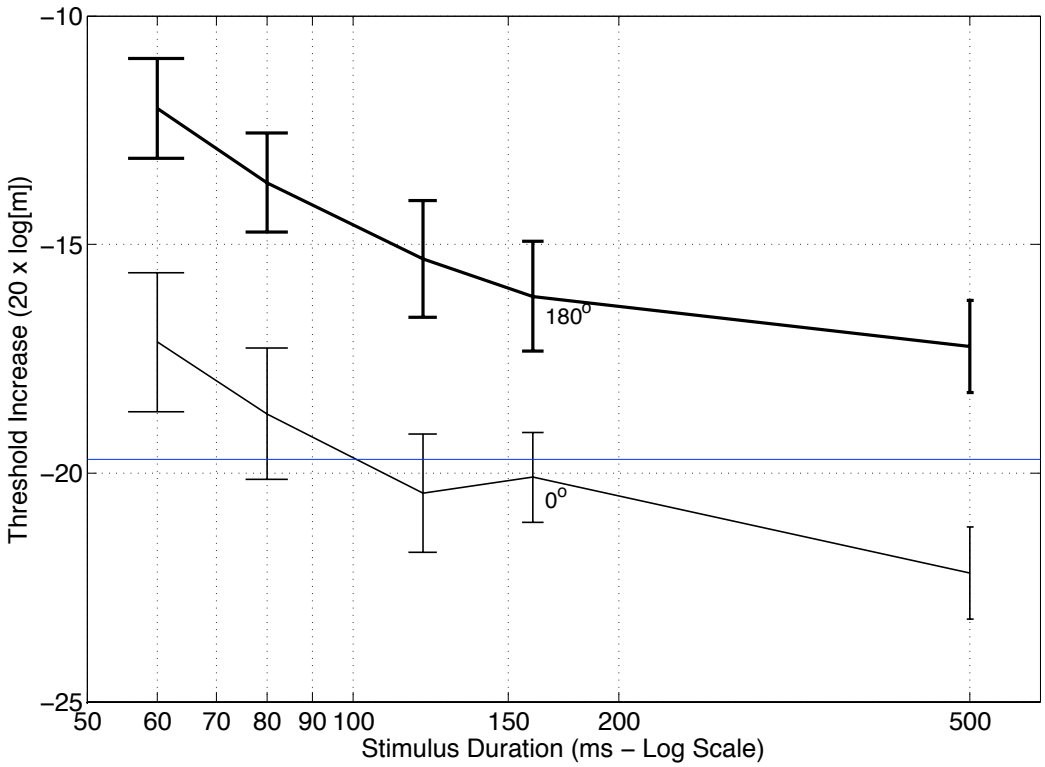

Figure 4.6: Subject normalised AM detection thresholds as a function of stimulus duration for phase conditions 0° (thin line) and 180° (thick line). Average baseline performance (for 500 ms stimuli) across all subjects shown as horizontal line. Error bars are 95% confidence intervals.

However, the fact that the threshold separation was maintained, suggests that the auditory system is able to detect modulation and to respond to changes in its phase, even at remarkably short durations.

4.7 Experiment 5: Vowel Environment in Engineered Real Speech Stimuli

Stimulus Interval Duration	80 ms	
Inter-Interval Gap	500 ms silence	
f_0 of Tone and Modulator	108 Hz	
Tone-to-Noise Ratio (TNR)	15 dB	
Phase (ϕ)	270°	
Noise Shape	Filtered white noise: [s]	
Vowel Environment	Natural and filtered /VFə/	(Scenario variable)
	V=*absent*, /*a*, *i*, *u*/ (Total 8)	
Vowel-to-Fricative Ratio (VFR)	15, 18, 25 and 35 dB	(Staircase variable)
Subjects	6 M, 6 F (Total 12)	
Threshold Runs per Estimate	1	

Table 4.5: Summary of conditions and parameters for psychoacoustic experiment 5.

From the literature discussed in 2.2.3 and as a direct consequence of the results of experiment 2, where AM detection was significantly impaired in the presence a loud, continuous tone, it was hypothesised that the environment in which the fricative burst was presented could affect AM detection. So far, the environment (i.e. the gap preceding and following each stimulus within the trial) has been either silence or a tone of the same amplitude as that which accompanies the frication noise.

In this experiment, engineered real-speech stimuli were used to present highly realistic fricatives in different vowel contexts. As well as the spectral properties (i.e., /ɑ,i,u/) and relative amplitude of the surrounding vowels, the effect of filtering to make them sound like non-speech sounds and that of removing the preceding vowel (simulating a word-initial fricative) were also investigated. In this way, it was intended to evaluate the suggestion from Section 2.2.3 that when stimuli are heard as speech, psychoacoustic thresholds may be affected, in this case AM detection.

12 subjects each completed a single threshold run for 4 vowel-to-frication amplitude ratios (VFR = {15 dB, 18 dB, 25 dB and 35 dB} under 8 individual conditions: 4 vowel conditions — no preceding vowel and /ɑ,i,u/, each either extracted from natural recordings or produced from LPC vowel-model filtering of a saw-tooth shaped waveform.

4.7.1 Method

The procedure for stimulus generation in Experiment 5 differs from that employed in Experiments 1–4 due to the fact that engineered real-speech stimuli are used. The procedure is described in detail in what follows and an outline of the signal processing appears as Figure 4.8 at the end of this section.

Original Recordings

Exemplar /VFə/ segments, $P(n)$ (where V=/ɑ,i,u/ and F=/z/; one segment for each vowel) were selected from the fluent-speech fricative corpus[4]. Waveforms for an example [azə] token are illustrated in Figure 4.7.

In order to eliminate possible bias due to varying fricative noise duration, a representative fricative duration of 80 ms was chosen (Pincas, 2004), and the corpus was scanned for examples with frication noise as close to this length as possible according to frication onset, F_{ON}, and offset, F_{OFF}, points identified in the acoustic study (Chapter 3).

Once short-listed, the final segments were chosen according to common f_0 during frication by manual comparison of pitch tracks. Whilst f_0 would have ideally been 125 Hz (as in experiments 1–4) in all cases, this was not possible and the final segments had $f_0 = 108 \text{ Hz} \pm 2 \text{ Hz}$.

For the chosen segments, frication onset and offset were then re-marked to coincide with up-going zero-crossings to ensure high quality splicing. This section of the original recording, $P(n_{F_{ON}\ldots F_{OFF}})$, consists of both frication noise and voicing. In order to construct a controlled stimuli, $P(n_{F_{ON}\ldots F_{OFF}})$ is rebuilt as follows.

Voicing During Frication

As the natural voicing found during frication is unpredictable (e.g., pitch jitter or glides, variable amplitude and devoicing) and hard to separate from the frication noise, the voicing component during frication, referred to as $a(n)$ in experiments 1–4, was replaced with a controlled voice recording.

Speaker PJ was asked to produce a sustained fricative with a stable pitch at the required f_0, and to gradually stop frication noise until only the weak voicing component remained; the resultant recording was then scanned manually for the most stable voicing period (in terms of amplitude and f_0). Voicing on and off points, V_{ON} and V_{OFF},

[4]Recordings from Speaker PJ were used as it was necessary that the subject be available for recording of the voicing segments and that the speaker be an able phonetician.

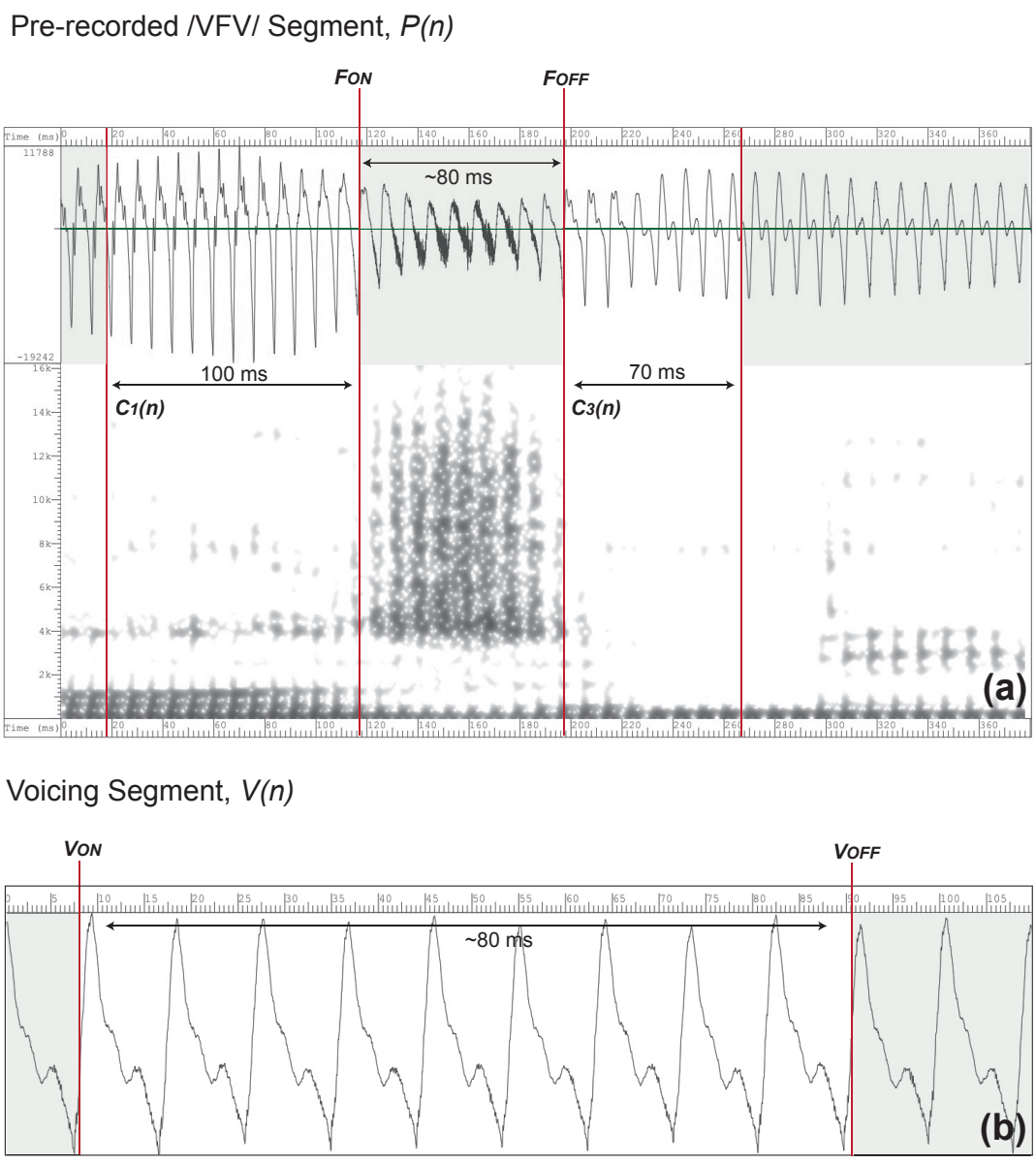

Figure 4.7: Stimulus construction in experiment 5. a) Waveform and spectrogram of [azə] recording used as basis for engineered stimulus. b) Waveform of [z] voicing recording used in construction of the stimuli. Vertical lines mark splicing points with according durations. Shaded areas are excluded from final stimulus. See text for full details.

were finally marked according to the requirement that splicing points be at up-going zero crossings and that duration be as close to 80 ms as possible. The result is $V(n)$: ∼80 ms of stable voicing, typical of the type found during VFs, but without frication noise, as illustrated in Figure 4.7(b). $V(n)$ is scaled in relation to the frication noise to attain a TNR of 15 dB to give $C_2(n)$.

Frication Noise

For the frication component, 80 ms of noise are generated and modulation controlled in the same manner as for experiments 1–4, described in Section 4.2.1. Spectral shaping for [z] is simulated using LPC filtering as in experiment 4 to give the final noise, $x(n)$.

Vowel Environment

The natural vowel environment was taken directly from the original recording, $P(n)$, as illustrated in Figure 4.7(a). For the segment preceding F_{ON}, 100 ms of vowel was cropped, giving $P(n_{F_{PRE}...F_{ON}})$, where F_{PRE} is the sample 100 ms (3200 samples at $f_s = 32$ kHz) prior to F_{ON}. Likewise, for the segment following frication offset, F_{OFF}, 70 ms of vowel are included, giving $P(n_{F_{OFF}...F_{POST}})$, where F_{POST} is the sample 70 ms (2240 samples at $f_s = 32$ kHz) after F_{OFF}.

Both vowel segments were then amplitude scaled to the required level before recombination and LPC filtering is applied if required[5], giving $C_1(n)$ and $C_3(n)$ for preceding and following vowel environments respectively.

Recombination

The stimulus is then rebuilt as the concatenation $C(n) = [C_1(n)\ C_2(n)\ C_3(n)]$ (omitting $C_1(n)$ in the case of the word-initial condition). To produce the final stimulus interval, the noise signal, $x(n)$, is added at $C(n_{F_{ON}})$ and the whole signal gated by the cosine ramp, giving $s(n)$.

Combination into trials and the experimental procedure for threshold estimation (3AFC, 2-up 1-down, variable step size) is the same as for experiments 1–4.

[5]For this 'filtered vowel' experimental condition, a 6th order LPC model of the vowel is used to filter a saw-tooth shaped wave, resulting in a vowel which has the same broad properties as natural vowel, but nonetheless does not sound speech-like.

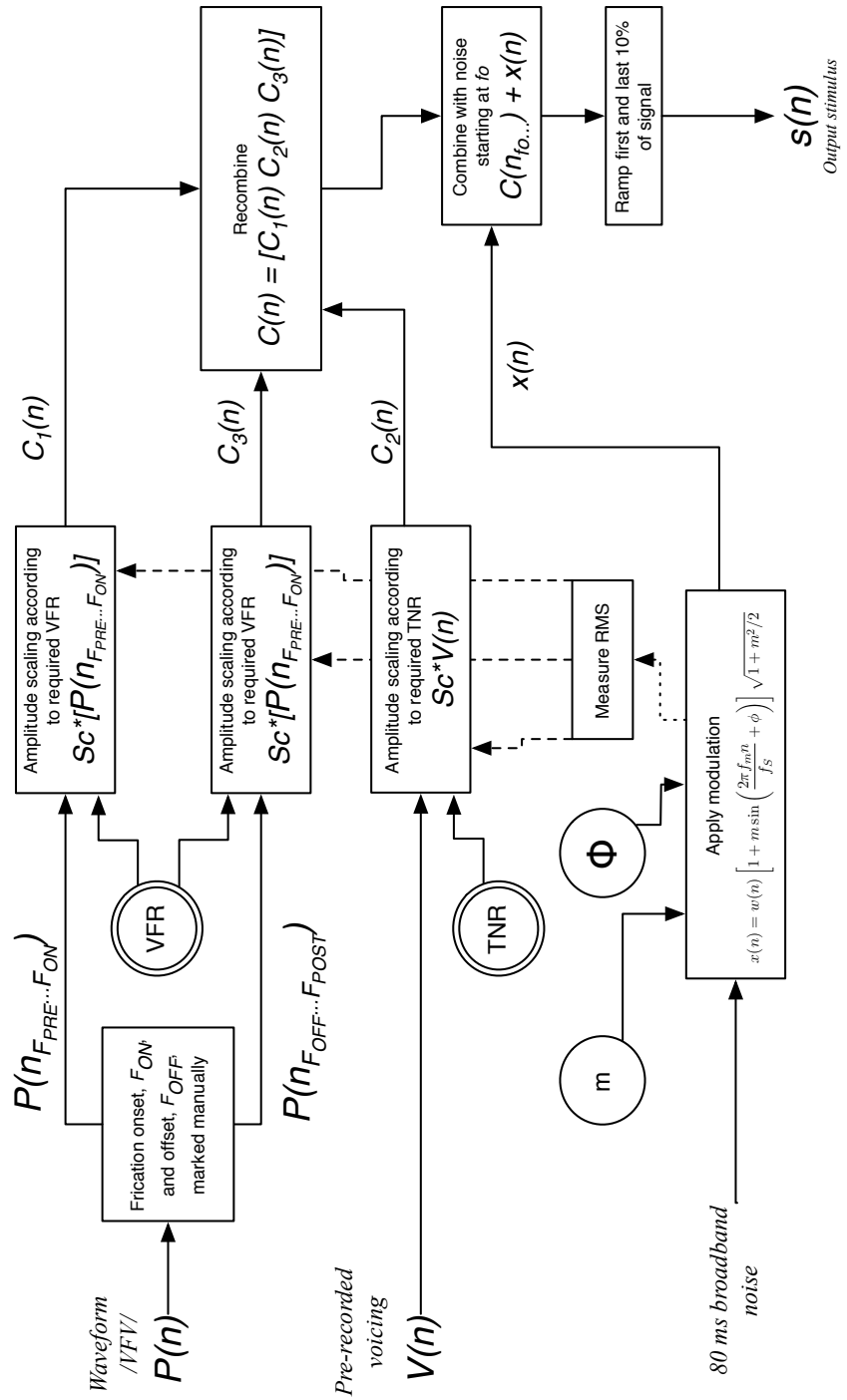

Figure 4.8: Construction of stimuli for Experiment 5. Rectangular blocks — processes; circular blocks — variables (single outline) and constants (double outline).

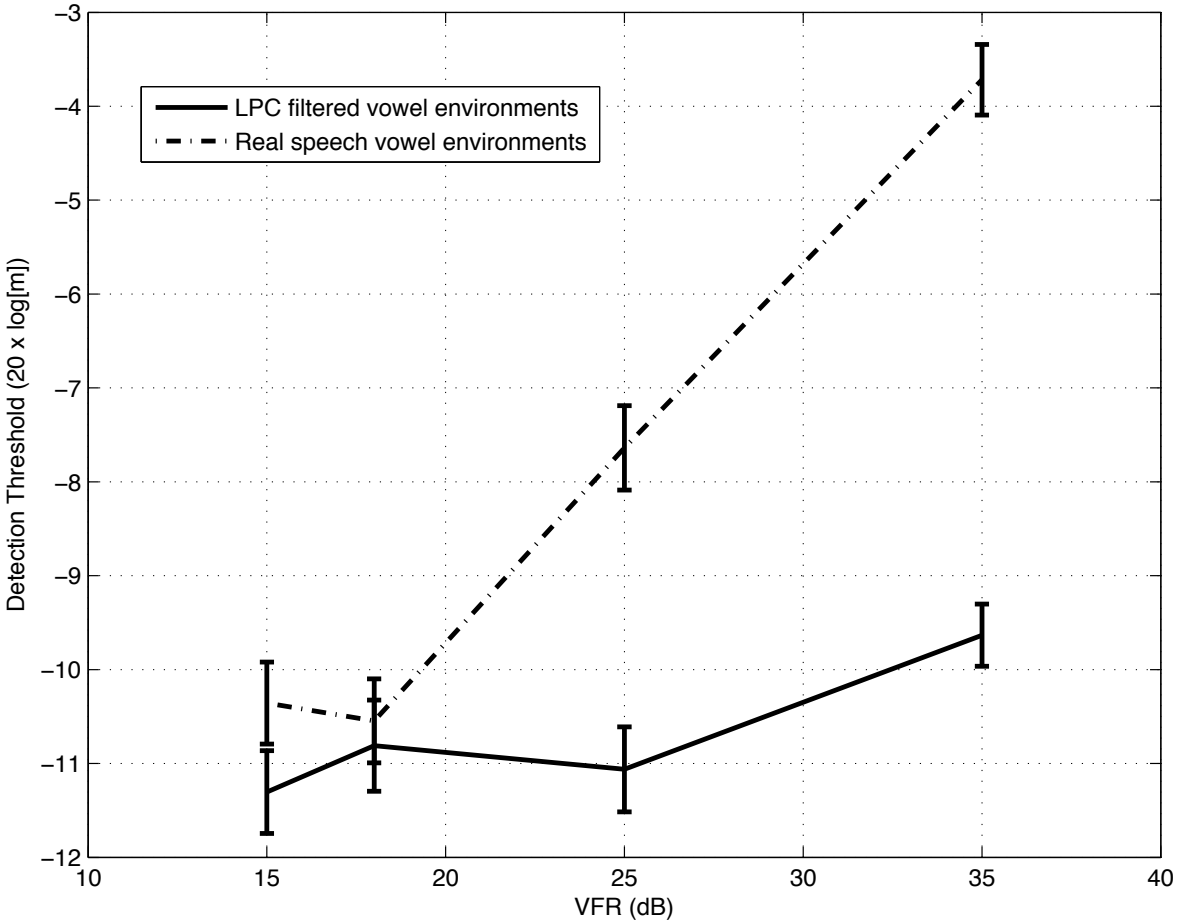

Figure 4.9: Relationship between VFR (dB) and AM detection threshold, m_d, for real vowel (dashed line) and LPC-filtered vowel environments (solid line). Data averaged over all vowel environments, word-initial/intervocalic contexts and speakers.

4.7.2 Results

A 5-way ANOVA (VFR, Vowel type, Vowel filtering, Initial vowel presence/absence, Subject) revealed main effect for VFR and Vowel filtering only, as well as an interaction between the two variables ($p < 0.05$). There appears to be no effect for vowel type or whether a preceding vowel is present or not (simulating word initial/intervocalic), so results are averaged over these conditions.

Figure 4.9 shows detection thresholds at all four VFs investigated for both real vowel and LPC simulated vowel conditions; results are averaged over vowel type, initial vowel condition and speaker. Moving from VFR=15 dB to 18 dB appears to cause no significant effect on detection threshold. At VFR=25 dB, the threshold increases by approximately 3 dB (impaired detection) in the case of the real vowel environment, but

there is no change for LPC vowels. At the highest VFR investigated, 35 dB, threshold for real vowels increases by a further 4 dB and there is a small but significant threshold rise over base performance of approximately 1 dB for filtered fricatives. At 35 dB VFR, threshold for AM detection is approximately 6 dB higher for real speech vowel environments than for filtered speech.

The effect of the amplitude of the vowel environment (VFR) and that for the speech/non-speech contrast can thus be summarised as follows. When stimuli are heard as non-speech, the amplitude of the vowel environment is of little importance, only impairing detection by a maximum of $\sim$1 dB for the loudest VFR (35 dB). On the other hand, when stimuli are heard as speech, the louder the vowel environment, the larger the impairment on AM detection: for VFR = 35 dB, there is a $\sim$6 dB higher threshold for speech stimuli compared to non-speech stimuli. This appears to accord with the two studies mentioned in 2.2.3 which both found impairment in detection/discrimination when speech stimuli were used.

4.8 General Discussion

4.8.1 Auditory Mechanisms

Our results suggests that a number of mechanisms may be responsible for the observed TNR and Phase effects.

For TNR in the range 10–30 dB, the effect of AM detection of the sinusoid depends strongly on its phase relationship with the noise. This suggests an 'additive' or 'multiplicative' masking relationship between the output of multiple cochlear filters across the spectral range and the low frequency tone, the explnation put forward by Wakefield and Viemeister (1985) in light of their similar results.

Above 30 dB, a different mechanism appears to take over, causing all thresholds to rise, irrespective of their phase. This has the effect of bringing 0° back to baseline performance and accelerating the rate of threshold increase for 180°.

One possibility for this latter mechanism is psychophysical masking of the noise by the tone, as introduced in Section 2.2.2. In the TMN paradigm presented in the current experiment, the pure tone at 125 Hz may mask a portion of the wide band noise carrier (16 kHz bandwidth), especially at higher TNRs, making the carrier less audible in the low-frequency range. Through a simulation of auditory filter output (Patterson-Holdsworth ERB filterbank model, (Slaney, 1993)), it can be shown that the upward-spread masking effect of the loudest tone (50 dB TNR) dissipates quickly. Figure 4.10 illustrates the ratio in dB of low-frequency tone to noise component amplitude (black

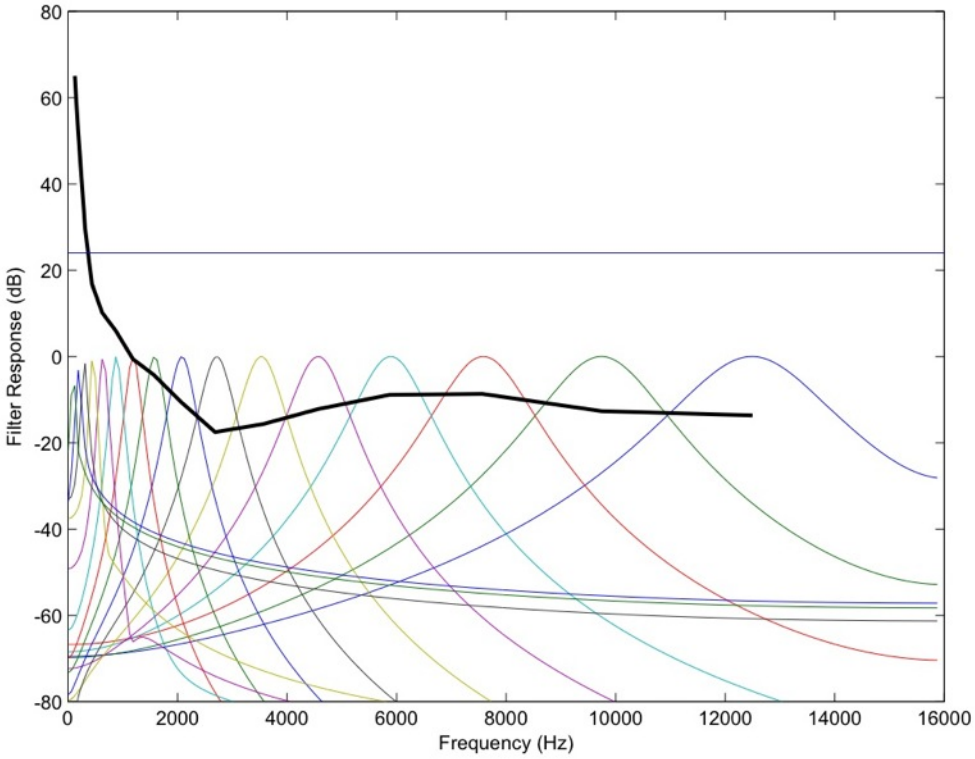

Figure 4.10: Power ratio output (black line) of auditory filters (coloured lines) simulated by the Patterson-Holdsworth ERB Filterbank Model. Ratio plotted at filter centres is for a 125 Hz tone to a 50 dB SPL 16 kHz BW broadband noise. Level of the tone is 50 dB above noise.

line) calculated at the output of simulated auditory filters (coloured lines) and plotted at filter centre frequencies. The effect of the tone has completely dissapeared by 2.5 kHz, suggesting that the masking capacity of even the loudest tone may be limited.

Furthermore, it does not follow that AM detection should be heavily affected by any reduced audibility of the carrier caused by masking. Previously reported data on AM detection and carrier level suggest that very quiet noise carriers can produce slightly raised thresholds (worse detection). Rodenburg (1972) found no differences in TMTFs for carriers at 20–60 dB, but Viemeister (1979) found a 3 dB difference between 0 dB and 20 dB carriers, which would account for raised thresholds at TNR $>$ 30 dB.

Further pyschacoustic measurement or more in-depth simulation work would be needed to establish whether TMN masking can be completely ruled out as an explanation for the effects obverved.

Another possibility is that the combination of the 50 dB SPL carrier and the loud-

est tones (resulting in an overal signal of ∼90–100 dB SPL) could produce distortion products in the testing equipment or middle ear, impeding detection of AM for those tone/noise combinations.

4.8.2 Is AM in VFs Perceptible?

Experiments 1–4 revealed that the acoustic complexities *internal* to VFs impair AM detection to only a limited degree. In experiment 1, the loudest tone (TNR = 50 dB) induced a highest detection threshold of -14 dB (m ≈ 0.20), a deterioration in performance of approximately 8 dB from baseline (no tone) performance. Only further 3 dB deterioration appears possible with the least favourable tone/modulation phase relationship (experiment 2, $\phi = 180°$, TNR = 50 dB, m_d = -11 dB [$m = 0.28$]). Furthermore, detection appears robust even faced with stimuli of duration matching the very shortest of VFs (experiment 4, $\phi = 180°$, TNR = 40 dB, $\lambda = 60$ ms, m_d = -12 dB [$m = 0.25$]). Of course, the most extreme values for each parameter (i.e., $\phi = 180°$, TNR = 50 dB, $\lambda = 60$ ms) will be mostly outside the range found for normal fricatives (or perhaps only represented in exceptional cases). Nonetheless, more acoustic work is need to establish precisely the range of values for TNR, phase and other parameters for VFs.

The worst case detection scenario is found for fricatives in a loud, natural vowel, intervocalic context (experiment 5, $\phi = 270°$, TNR = 15 dB, $\lambda = 80$ ms, VFR = 35 dB, Natural vowel context, m_d = -4 dB [$m = 0.63$]), confirming not only that AM detection is strongly affected by the relative volume of the surrounding sounds, as might be predicted from the literature on nonsimultaneous masking, but also that a natural speech context (as opposed to sounds that are heard as non-speech) is significantly more effective at impairing detection. Although the latter finding has important implications for research bridging psychoacoustic and speech perecption, it should be noted that VFR = 35 dB is an extreme value that is very unlikely to occur in real fricatives, although, again, measurements of VFR in the literature are lacking.

Assuming a VFR ranging from 15 to 25 dB and the possibility of a 3 dB performance deterioration for TNR, phase and duration noted above, a realistic estimate of AM detection thresholds for most VF conditions is -11 dB $\leq m_d \geq$ - 5 dB (or $0.28 \leq m \geq 0.56$). Compare these values to the results of the acoustic study: peak modulation for any fricative — 0.65, mean peak modulation — 0.5 and overall mean modulation — 0.35. These values should be interpreted with caution as m changes throughout a fricative in conjunction with voicing strength, as previously noted. Thus peak modulation within a fricative is observed where voicing is strongest, mainly at frication onset. In conclusion, then, AM should be perceptible, at least through a part of frication, for a wide range of fricatives, although probably not in all cases.

4.9 Summary

A series of five AM detection threshold measurement experiments investigated the detectability of AM in stimuli designed to reproduce the acoustic attributes of voiced fricatives. Stimuli ranged from simple noise-plus-tone simulations to engineered real-speech stimuli that were heard as realistic fricatives. The experiments were carried out with the aim of establishing what acoustic properties of VFs, if any, might impair AM detection to the extent where it becomes unavailable as a potential cue to the perception of voicing in fricatives.

All paramaters tested (with the exception of spectral shaping of the noise) had the potential to impair AM detection: increasing amplitude of the simultaneous tone (representing voicing), decreasing duration of stimulus and increasing amplitude of the vowel environment. It was also shown that when VFs are presented in realistic speech environments rather than synthetic tones, detection is impaired significantly.

The effect of the phase relationship between modulating signal and simultaneous tone was more complex, with detection *enhancement* possible as well as impairment. A 180° difference in phase gave the largest difference between enhancement and impairement, although the magnitude of the former effect was minor compared to the latter.

Overall, it is esimated that detection thresholds for AM in VF's ranges approximately -11 dB$\leq m_d \geq$- 5 dB (or $0.28 \leq m \geq$ 0.56). Thus, despite the complex acoustic characteristics of VFs leading to impairment of AM detection in most cases, it is concluded that the magnitude of the effects was small enough for detection to remain robust in real VF's.

Chapter 5

Cue-trading Experiment

5.1 Introduction

Having established that AM perception for signals resembling VFs is robust, experiments reported in this chapter investigated how listeners might use it as a cue to fricative voicing by examining whether AM trades with other cues to the voicing distinction that have been suggested in the phonetics literature (see Section 2.3 for full review of the literature relating to these cues), i.e.:

1. *Voicing, presence during frication*
2. *Voicing/frication, overlap duration at onset or offset*
3. *Voicing, amplitude during frication (c.f. TNR)*
4. *Frication, duration*
5. *Formant transition, duration prior to frication onset (c.f. PP)*
6. *Preceding vowel, duration*

AM of frication noise has not previously been seriously considered in this context. This experiment thus combines the above variables with the following novel cues:

1. *Amplitude modulation of frication, presence*
2. *Amplitude modulation of frication, depth (m)*
3. *Amplitude modulation of frication, phase (ϕ)*

Voicing presence/amplitude and *fricative duration* were initially identified as good candidates for the construction of a continuum as they are well proven in the literature and straightforward to manipulate acoustically (Denes, 1955; Cole and Cooper, 1975; Stevens et al., 1992).

However, after various initial attempts at stimulus engineering, it was not possible to create a continuum where the extremes were judged as 100% voiced/voiceless (a useful stimulus quality control, see Soli (1982)) without varying other parameters. Recall the argument from Soli (1982) (see Section 2.3.3) that stimuli varying fricative duration or voicing amplitude only "do not contain the strong vowel structure and duration cues for voicing that are present in naturally produced vocalic stimuli." It is thus unsurprising that a continuum based on these parameters alone should have been perceptually unsuccessful.

It was subsequently found that *formant transition* produced the most realistic voicing continuum, with the extremes being correctly interpreted close to 100% of the time (see 'Results'). This cue has previously been confirmed in two important studies (Soli, 1982; Stevens et al., 1992), the findings of which have implications for the research discussed in Section 2.3.3 that posits *frication duration* as a primary perceptual cue, along with the traditional view that it is acoustic *voicing presence* that specifies phonological voicing.

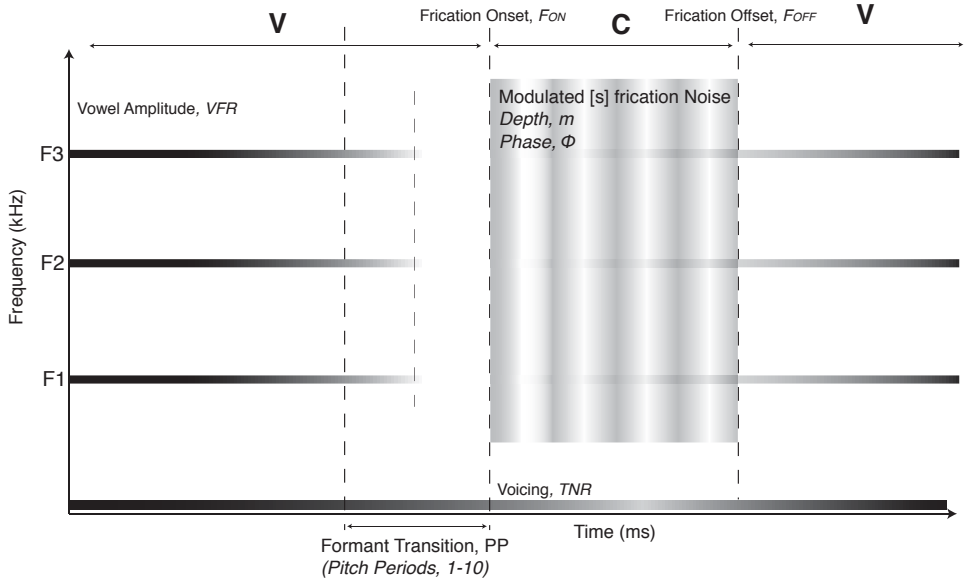

Figure 5.1: Schematic illustration of /VCV/ stimulus.

Using a natural [z] fricative embedded in a nonsense word and phrase as a basis for the engineered stimuli, voicing responses were obtained along a formant-transition-duration continuum measured in voicing pitch periods.

Figure 5.1 illustrates the general format of stimuli as a schematic spectrogram. Notice how the formants fade and die out prior to the onset of modulated frication. For any specified transition duration, the first half (e.g., 4 PP for an 8 PP transition; illustrated by the intermediate dashed line in Fig. 5.1) is a formant energy rampdown. The second half is characterised by a complete absence of formant energy[1].

Six transition durations were tested PP = 0, 2, 4, 6, 8, 10). The maximum PP = 10 transition was decided on the basis of informal listening tests. For transitions longer than 10 PP it was observed that no further increase in the strength of the percept of voicelessness was possible; instead, stimuli began to sound unnatural.

In the main experiment, *AM absence or presence* was tested. The AM depth for the modulation-present condition was determined by the minimum level required for widespread detection according to the results of psychoacoustic experiment 5 (Section 4.7). At VFR = 15 dB (the value used in this experiment; recall that VFR refers to the level of the surrounding vowel), experiment 5 suggests a detection threshold of approximately -10.5 dB ($m = 0.3$) which is the AM depth used in this experiment. Keeping AM as low as possible (but within the detectible range) allows for maximum generalisation of results: $m = 0.3$ represents a conservative amount of modulation across fricatives.

In an extension, *AM depth* was varied ($m = 0.1$, 0.5, 0.7) along with *phase* (0° and 180°). AM values were chosen to bring modulation depth into the highly detectible range. Psychoacoustic experiment 5 showed that even at an exaggerated VFR level of 35 dB, AM with $m = 0.63$ (-4 dB) is detectible. Phase values represent the two extremes of the effect observed in experiment 2 (Section 4.4).

In addition to *formant transition*, *AM absence or presence*, *AM depth* and *phase*, the *presence/amplitude of voicing* (TNR = -∞, -10, 0 and 10 dB) was also included as a cue parameter. Traditionally, presence of voicing is thought of as the primary cue to the phonological voicing distinction, thus it was of interest to determine how any trading relations discovered for the main cue parameters patterned across voicing (TNR) conditions that could be conflicting (e.g., strongly suggesting voiceless where the combination of other cues strongly suggests voiced), neutral or supporting.

The final parameter, low-frequency masking, was included to replicate possible real scenarios (environmental noise, poor quality telecommunications systems) where the low-frequency glottal vibration cue to voicing may be unavailable to the listener. In the case of TNR = $-\infty$ the glottal vibration cue is also unavailable, but in this case

[1] The weak aspiration noise in the higher frequency bands is, however, left intact. This pre-aspiration noise is characteristic of voiceless fricatives (as well as other consonants) and its presence is essential for natural sounding stimuli.

it is overtly absent and thus the hypothesis is that mostly unvoiced responses will be elicited. In the case of low-frequency masking, the listener does not know whether the glottal vibration is actually absent or simply masked by the noise and might thus be expected to make more recourse to secondary cues, such as AM.

Categorical perception experiments in phonetics involve the manipulation of a known cue along a continuum where one extreme elicits a particular phonetic category response (with 100% of responses, or close to) and the other extreme elicits the opposite response. For non-binary categories, such as place of articulation, interim points along the continuum may elicit other categories. In cue-trading experiments, a shift in the response function is produced by varying a secondary cue of interest (Repp, 1982). This shift essentially represents the concept that an identical subjective response (for example, 70% 'voiced' responses) can be achieved with entirely different combinations of cue parameters.

The research question motivating this experiment is whether AM can act as a cue to voicing at all, and if so, to what extent and in which ways does it trade with other cues. It is most often assumed that the largest cue effect for a parameter of interest (in this case modulation) is found when values of other parameters are 'neutralised', i.e., their setting does not strongly represent either (or any) phonetic category. If *AM presence* does cue voicing, it is therefore hypothesised that the largest effect will be at intermediate values of the formant transition continuum. It is harder, however, to predict how *AM presence* might trade with *TNR* as the latter is presumably highly categorical (i.e., exhibiting a very steep gradient in the CP characteristic, admitting few values close to 50% 'voiced' responses): either a stimulus is consistently heard as 'voiced' or 'unvoiced', possibly limiting the potential effect of *AM presence*.

It is also difficult to predict how deeper modulation and phase manipulations might affect access to AM as a cue. The psychoacoustic study suggested that modulation may be only marginally detectible (or even undetectable) in some VFs. Therefore increasing its detectability through larger depth or more salient phase would increase the proportion of occasions on which it could be detected and potentially used by listeners. Assuming that AM *can* act as a cue, increased detection would presumably lead to increased 'voiced' responses. It is not clear, however, whether increasing AM depth once detection has already been achieved (i.e. simply augmenting the perceptual salience) would go any way to increased 'voiced' responses. The experiment should thus determine whether AM detection produces the maximal effect on voicing judgements, or whether more perceptually salient modulation will induce more voicing responses in a manner akin to other known cue continua.

Section 5.2 presents the detailed method for the generation of the stimuli described above as well as the experimental procedure used for results collection. Section 5.3

presents results and discussion from the main experiment, with Transition, Modulation Presence, TNR and Masker Presence as variables, and extension experiment testing Modulation Depth, m, and Phase, ϕ.

5.2 Method

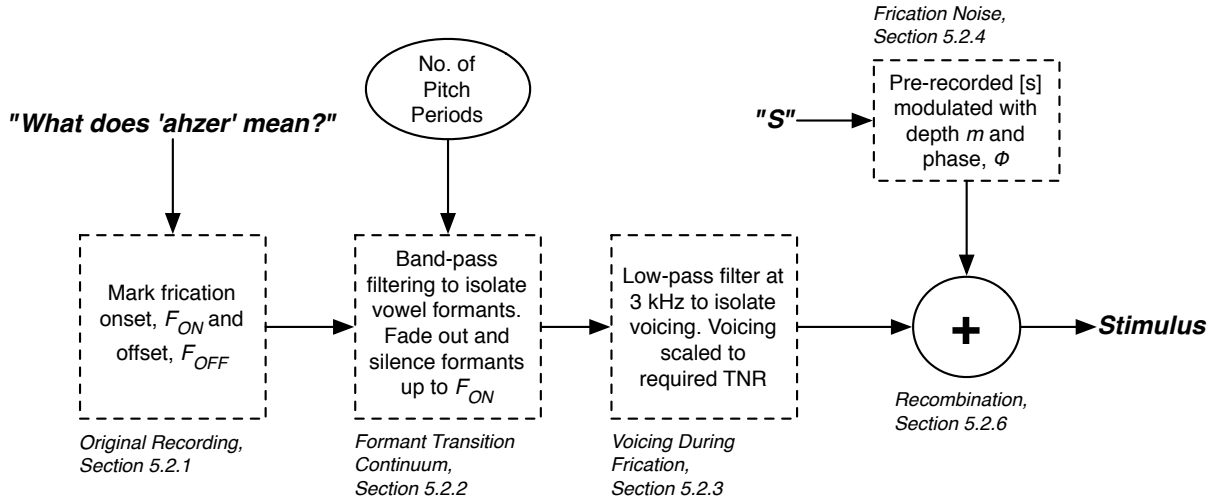

Figure 5.2: Overview of signal processing for stimulus generation.

An overview of stimulus production is presented as Figure 5.2. A recorded [z] and vowel context are deconstructed into three sections (as shown in Figs. 5.1 and 5.4) that are manipulated and finally recombined. The first section is up to the start of frication. Here, the formant transition into the fricative is faded as required by band-filtering. The second section is the voiced fricative, which is constituted of voicing obtained by low-pass filtering the section itself, and [s]-noise from a second recording which is modulated according to the experimental condition. The third section is the vowel environment following frication offset. The amplitude of the first and third sections is scaled to a specified VFR (vowel-to-frication ratio, see psychoacoustic experiment 5) before recombination into the final stimulus.

Signal processing operations for stimulus generation are described and illustrated in the following and summarised in Figure 5.8 at the end of this section.

5.2.1 Original Recording

An exemplar recording of a token of [z] within the nonsense word /azə/ and embedded in the phrase "What does /azə/ mean?" was selected from the fluent-speech corpus recorded for the acoustic study (designated $P(n)$ in Fig. 5.8). The token was selected on the basis of stable f_0 during voicing to ensure that modulation phase could be accurately constructed. A set of tokens from the corpus was selected where f_0 during

the vowel prior to frication onset was 125 Hz +/- 20 Hz. Manual comparison of pitch tracks during voicing were then used to identify the token with most stable f_0 during voicing frication. The f_0 of the selected token was constant at 125 Hz throughout frication.

Frication onset, F_{ON}, and offset, F_{OFF}, adjusted to the nearest up-going zero-crossing were also manually annotated.

5.2.2 The Formant-Transition Continuum

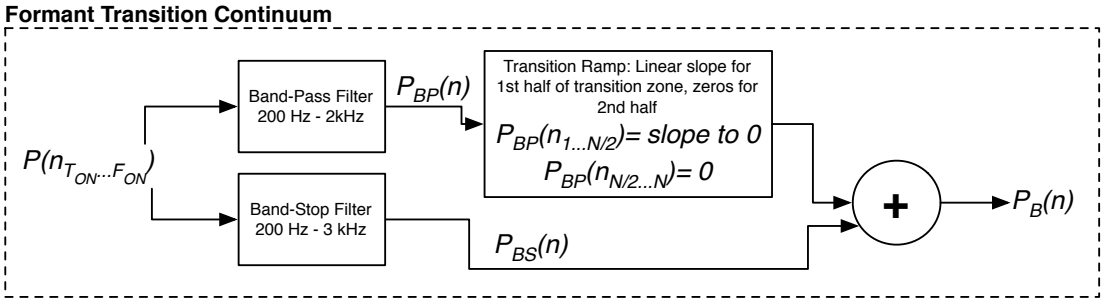

Figure 5.3: Summary of signal processing for generation of the formant-transition continuum.

In order to build the formant transition continuum, a principled method of describing its duration relative to a fixed end-point was required. Transition duration was thus specified as number of glottal pitch periods (PP = 0–10) prior to frication onset. This method was preferred to specifying the duration in millisecond time increments as it allowed for splicing at zero crossing points, avoiding signal artefacts on recombination.

Zero crossing points, T_{ON}, for each of the 10 pitch periods were manually annotated and input to the algorithm to substitute for PP.

Copies of the signal from the transition starting point to the frication onset, $P(n_{T_{ON}...F_{ON}})$, are band-pass (200 Hz – 2 kHz, 2-pole) and band-stop (200 Hz – 3 kHz, 2-pole) filtered, giving $P_{BP}(n)$ and $P_{BS}(n)$ respectively. Approximately the first three formants (most perceptually important) are isolated in $P_{BP}(n)$ and the voicing transition is created by manipulating this signal. Transition zones consisted of two phases, each lasting half the total transition duration: first, a linear ramp down of formants; second, total absence of formant energy, as illustrated in Figure 5.4 and described by Equation 5.1:

$$\overline{P}_{BP}(n) = \begin{cases} w(n)[P_{BP}(n_{1...N/2})] & \text{if } n \leq \frac{N}{2} \\ 0 & \text{if } n > \frac{N}{2}, \end{cases} \tag{5.1}$$

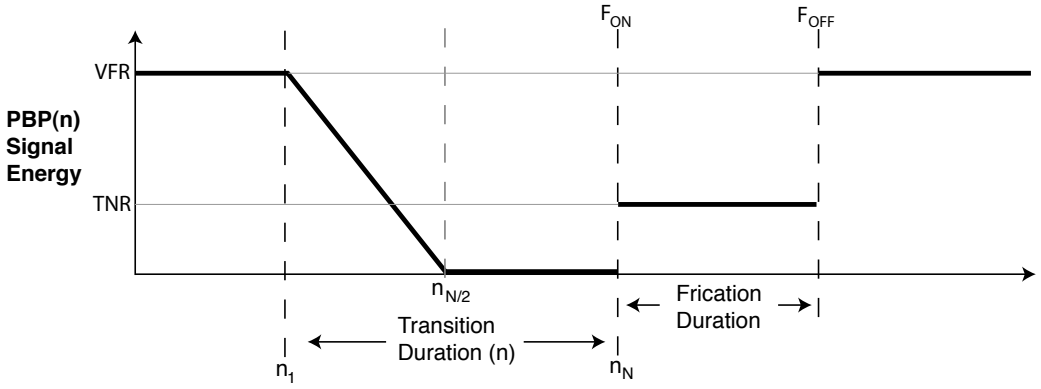

Figure 5.4: Schematic illustration of the energy rampdown in generation of the formant-transition continuum.

where $w(n)$ is a linear ramp $[1 \ldots 0]$ of length $N/2$. The ramped formant transition, $\overline{P}_{BP}(n)$, is then recombined with the low-frequency voicing and higher frequency components, $P_{BS}(n)$, to give $P_B(n)$.

5.2.3 Voicing During Frication

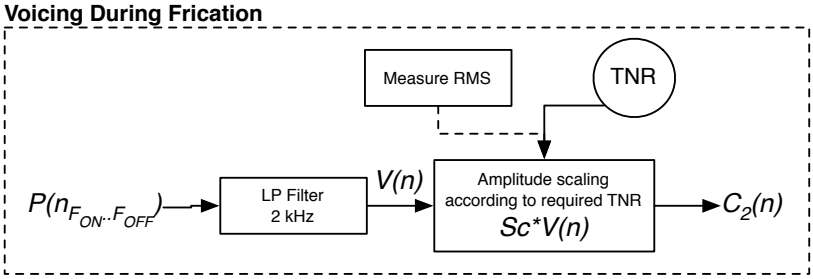

Figure 5.5: Summary of signal processing for generation of the voicing component. The scaling factor, Sc, is defined in Equation 5.2.

Natural voicing during the frication is created by low-pass filtering (3 kHz, 4-pole) the portion of the exemplar token between frication onset and offset, $P(n_{F_{ON} \ldots F_{OFF}})$, to remove frication. The resulting voicing signal, $V(n)$, is then scaled according to the required TNR as

$$C_2(n) = V(n)(\frac{F_{rms}[10^{\frac{TNR}{20}}]}{V_{rms}}), \tag{5.2}$$

where F_{rms} is the frication noise RMS, V_{rms} is the RMS of $V(n)$, and $C_2(n)$ is the resulting scaled voicing signal ready to be incorporated into the final stimuli.

5.2.4 Frication Noise

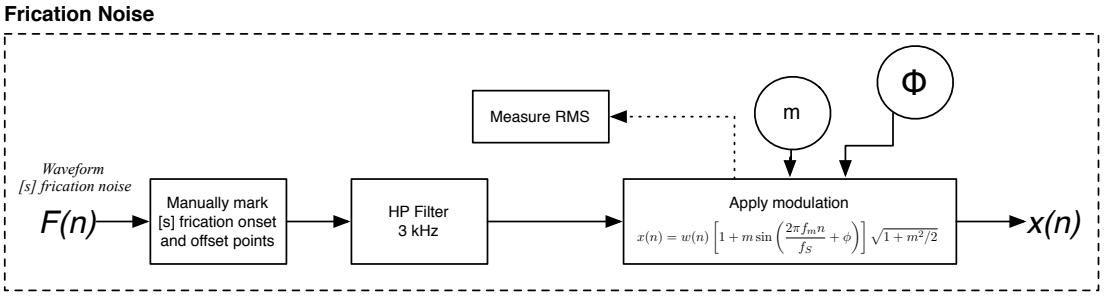

Figure 5.6: Summary of signal processing for generation of the modulated noise component.

An exemplar token of [s] noise of the required duration was selected from the fluent-speech corpus and the word and sentence context removed. The cropped waveform, $F(n)$, was then high-pass filtered to ensure complete absence of low-frequency components. Modulation is applied as required by the experimental condition in the same manner as for psychoacoustic experiments 1–5 (Equation 4.2) giving $x(n)$.

5.2.5 Vowel Environment

The sections of the recording prior to frication onset, F_{ON}, and following frication offset, F_{OFF}, are amplitude scaled relative to F_{rms} and the required VFR (see Equation 5.2) to give $C_1(n)$ and $C_3(n)$. As VFR is meant to refer specifically to the vowel environment surrounding frication, only 100 ms of vowel preceding frication and 70 ms following frication are included in the RMS measurement for VFR calculation, although the resultant amplitude scaling is obviously applied to the entire recording before and after frication to avoid abrupt changes in amplitude.

5.2.6 Recombination

The preliminary stimulus, $C(n)$, is reconstructed as $[C_1(n)\ C_2(n)\ C_3(n)]$. The modulated noise signal, $x(n)$, is then added at $C(n_{F_{ON}})$.

As trial intervals are separated by silence, the whole stimulus is gated to avoid sudden and jolting onsets and offsets of sound. The initial 10% rise and final 10% fall of

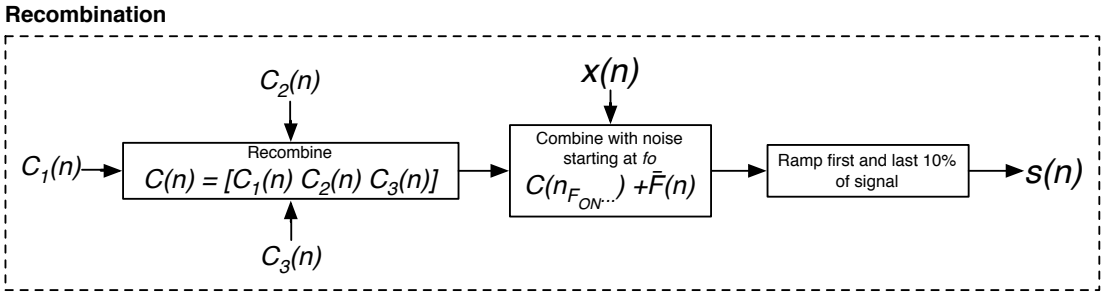

Figure 5.7: Summary of signal processing for recombination of signals into the final stimulus.

$C(n)$ were gated by a raised cosine ramp, $h(n) = \frac{1}{2}\left(1 - \cos\frac{\pi n}{N}\right)$, where N is the ramp duration.

5.2.7 Low-Frequency Masking

If low-frequency masking is called for by the experimental condition, a waveform of low-pass filtered (2 kHz, 4-pole) rumble is added to $C(n)$ at 10 dB above the amplitude of voicing during the fricative (TNR).

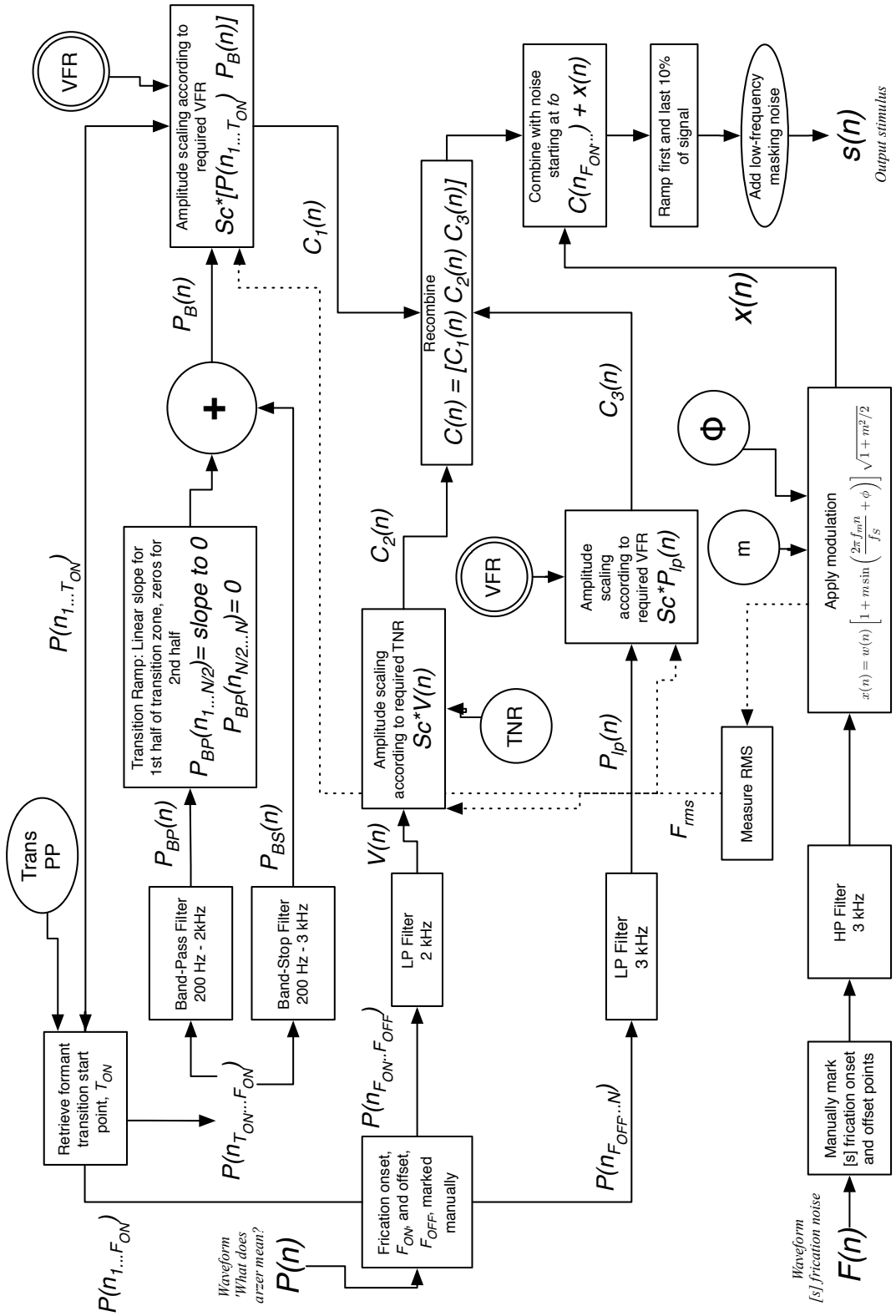

Figure 5.8: Construction of stimuli for cue-trading experiment. Rectangular blocks — processes; circular blocks — variables (single outline) and constants (double outline).

5.2.8 Experimental Procedure

Parameter	Exp. 1	Exp. 2
f_0	125 Hz	125 Hz
VFR	15 dB	15 dB
TNR	-∞, -10, 0, 10 dB	0 dB
m	0.3	0.1, 0.5, 0.7
ϕ	0°	0°, 180°
Transition Gap (PP)	0, 2, 4, 6, 8, 10	3, 5, 7
LF Masking	Yes/No	No
Subjects	10	6

Table 5.1: Parameters, variables and constants for cue-trading experiments.

Stimuli for all unique experimental parameter combinations (summarised in Table 5.1) were pre-generated at $f_s = 32$ kHz and stored on and presented from computer disk. Using the same equipment set-up and calibration procedure as for the psychoacoustic experiments (described in Section 4.2.3), subjects are presented with the stimuli containing the phrase "What does /VCV/ mean?" and respond according to whether they judge the word to contain a voiced or voiceless fricative (i.e., whether they hear 'arser' or 'arzer'). Each combination of parameters was presented a total of 10 times. Order of presentation of parameter combinations was randomised.

5.3 Results and Discussion

Results for the unmasked (thin lines) and masked (thick lines) conditions for the main experiment are shown in Figure 5.11 ('Modulation' section below). Percentages of voiced responses (/z/) for each unique combination of stimuli conditions are across all speakers, corresponding to 100 presentations in total (10 subjects × 10 presentations each).

Following the example of Soli (1982), it was confirmed that canonical voiced and voiceless fricatives were heard correctly. Stimuli with no voicing (top left panel) or modulation (crosses, black lines) and maximal transition gap (10 pitch periods) were heard as voiceless ∼95% of the time. Likewise, in the loudest voicing condition (10 dB, bottom right panel), stimuli with no transition gap (0 pitch periods) and modulation (red line, circles) were heard as voiced in ∼95% of cases. So, it can be concluded that engineered stimuli maintained their quality and phonetic properties.

5.3.1 Unmasked Stimuli

Transition

Transition Gap is a strong cue to voicing under the majority of conditions. Figure 5.9 shows the effect for Transition averaged over all four TNR conditions with confidence intervals computed across subjects' results. Voiced responses to stimuli across a 10 pitch-period transition gap (∼80 ms) continuum span the 50% crossover point and were 58% less (unmodulated stimuli) and 49% less (modulated stimuli) for full, 10 PP Transition stimuli than for stimuli with no gap (i.e., the natural transition for /z/).

This effect is comparable to that reported in Soli (1982), who compared voiced responses for natural voiceless and voiced transition types, based on /jus/ and /juz/, and estimated 30–40% more voiced responses for the /juz/ transition. However, the results are in broad agreement that the structure of the formant transition from the vowel preceding frication is an important cue to fricative voicing.

TNR

The effect for TNR is illustrated in Figure 5.10 where voiced responses at PP = 5 are shown for unmodulated (blue) and modulated (red) stimuli. In both cases, percentage voiced responses rise as TNR increases; the difference in responses between the smallest and largest TNR is 25%– 35%, with the larger difference for unmodulated stimuli.

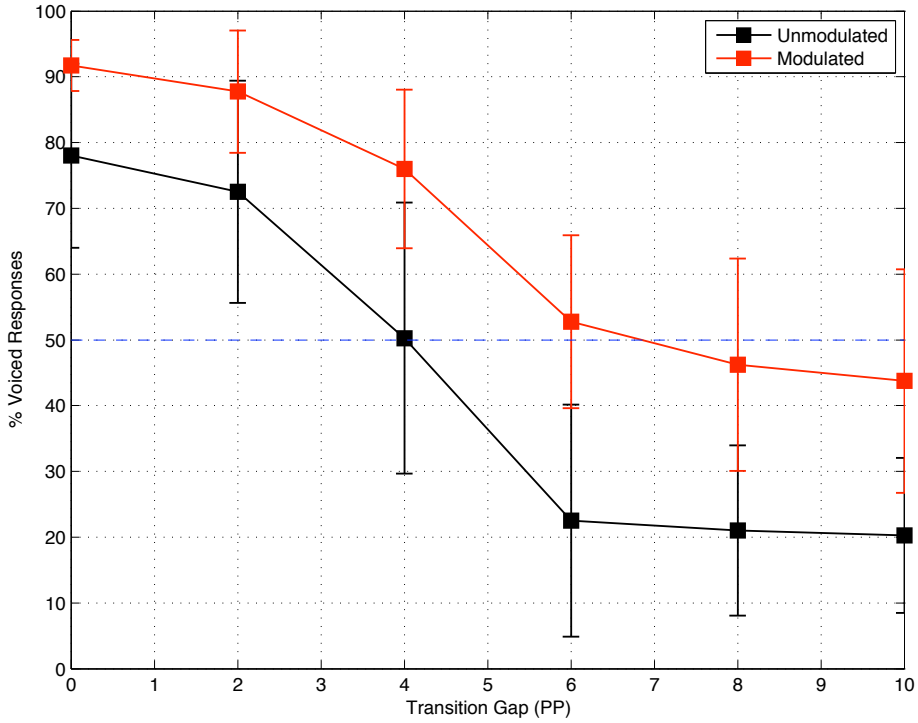

Figure 5.9: Percentage 'voiced' responses at points along a transition continuum measured in pitch periods (PP) for modulated (red) and unmodulated (black) conditions averaged over all TNR conditions. Error bars show 95% CI intervals across subjects' results.

Modulation

The effect for modulation is apparent in both Figure 5.9 and Figure 5.10. Where results are averaged over TNR (Fig. 5.9), modulated stimuli produce consistently more voiced responses than unmodulated stimuli although the difference is not statistically significant at the $p < 0.05$ level due to wide variation in results for different TNRs that will be discussed below. For the results across TNR at the transition mid-point (Fig. 5.10), the difference in voiced responses between unmodulated and modulated stimuli is observed to decrease as TNR increases; at TNR = 10 dB, the Modulation effect is no longer significant.

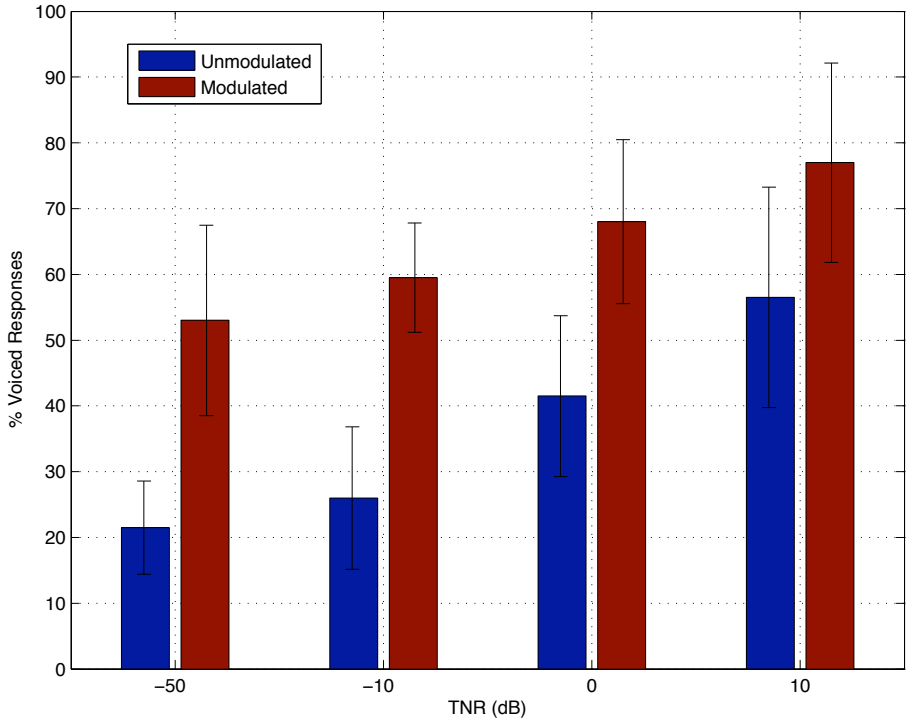

Figure 5.10: Percentage 'voiced' responses for intermediate transition-gap stimuli (PP = 5) as a function of TNR for modulated (red) and unmodulated (blue) conditions. Error bars show 95% CI intervals.

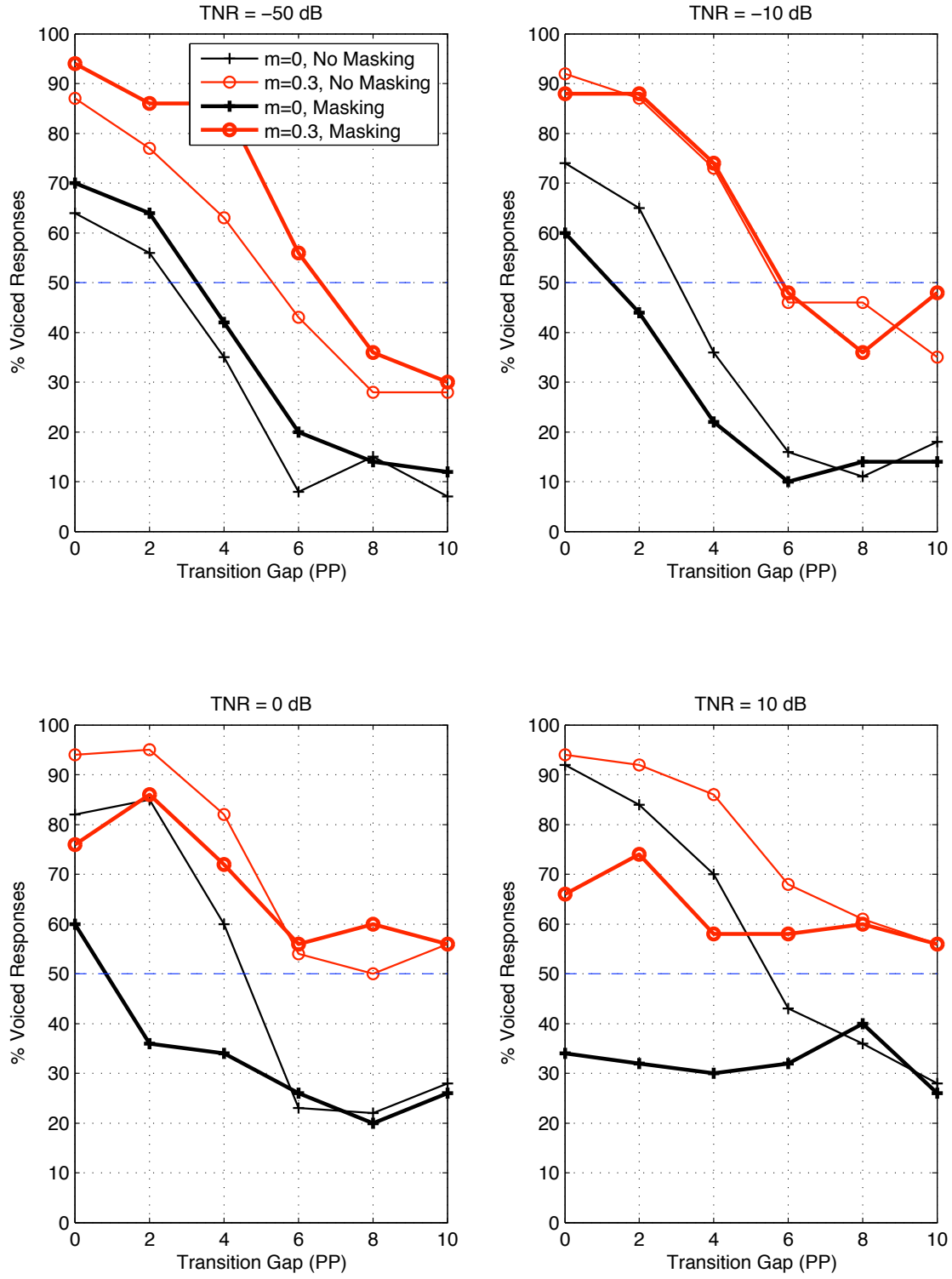

Figure 5.11: Percentage 'voiced' responses at points along a transition continuum measured in pitch periods (PP) for stimuli with (thick lines) and without (thin lines) low-frequency masking and for modulated (red lines, circles) and unmodulated (black lines, crosses) conditions. Frames show results for TNRs (top left: no tone; top right: -10 dB; bottom left: 0 dB; bottom right: 10 dB).

Figure 5.11 presents results for all four TNRs individually, allowing observation of the Modulation effect for varying Transition and TNR conditions. The size of the perceptual effect of Modulation can be gauged from the vertical distance between unmodulated (thin black) and modulated (thin red) functions for each TNR condition at points along the transition continuum. Where voicing is absent or weak (top panels), modulation causes a 20%–30% increase in voiced responses which is approximately uniform across the transition continuum. For louder TNRs (bottom panels), the effect of Modulation appears to start smaller for shorter Transition gaps (at PP = 0, approx. 10% difference for TNR = 0 dB; no difference for TNR =10 dB) and increase up to 30% for the longest Transition.

Statistical significance for the effect of modulation is gauged using a two-sample χ^2 test (Test of Association) with Yates' correction for continuity (defined in Appendix C.2)[2]. For each combination of TNR and Transition gap, the voiced responses for modulated and unmodulated stimuli are tested as a 2×2 contingency table. For example, for TNR = -50 dB and PP = 10:

	Unvoiced	Voiced
Modulated	93	7
Unmodulated	72	28

In this case, Yates' $\chi^2 = 13.85$ which is significant at the $p < 0.00$ level. Table 5.2 shows p-values to 2 d.p. for χ^2 tests comparing modulated and unmodulated responses at all combinations of TNR and Transition Gap. Note that in contrast to post-hoc testing following ANOVA analysis, there is no accepted procedure for correcting for multiple χ^2 tests (Camargo et al., 2008), so values are unadjusted.

It can be seen that the effect for modulation is significant to the $p < 0.05$ level in all but the PP = 0 and 2 Transition Gap (0 and ∼16 ms) cases for the TNR = 10 dB condition (confirming the previously made observation that the effect of Modulation is smaller for shorter Transition Gaps for TNR = 0 and 10 dB).

The most notable result was thus obtained with the more strongly voiced stimuli (similar to real voiced fricatives). Loud voicing combined with a transition strongly suggestive of a voiced fricative was enough to elicit close to 100% voicing judgements and largely eliminate the effect of modulation. However, as the transition gap lengthens,

[2]The Yates' χ^2 test is generally used when at least one cell of the contingency table has en expected frequency less than 5 (Yates, 1934), although some sources recommend 10. Application of the Yates' correction leads to a more conservative p-value than a normal χ^2 test. In the present data, some expected frequencies are less than 5 and 10, but the Yates' correction is nonetheless applied throughout, leading to estimations of p that are generally more conservative.

the size of the effect caused by modulation gets larger. Thus, even in the presence of strong voicing that would typically be expected to dominate, modulation elicits more voiced responses if the transition cue is neutral or contradicts the voicing cue.

The horizontal distance between modulated and unmodulated functions illustrates the cue-trading relation between Transition Gap and Modulation directly. Consider that inclusion of modulation produces the same perceptual result (increased percentage of voiced responses) as shortening the transition.

For intermediate transition stages where the changeover from voiceless to voiced responses is sharpest (PP = 4–6, ∼32–48 ms), equivalent transition reduction is approximately 1.5–2.5 PP (∼12–20 ms) depending on TNR. In other words, a 30% increase in voiced responses induced by modulation could also have been elicited with an unmodulated stimulus with a transition gap 2 pitch period shorter.

Across TNRs, the maximum range of Modulation effect appears to be equivalent to approximately 1.5 PP (∼12 ms) for all but the shortest, 4 PP, transition where the range is expanded to 3.5 PP (∼28 ms) with a notably larger Modulation effect for no voicing and weak voicing (TNR = -10 dB) cases.

It has been suggested that the emphasis on unnatural, 'neutralised' stimuli is a weakness of cue-trading experiments. According to this argument, cue-trading experiments typically find significant effects for the stimuli of interest only where supposed 'primary' cues have been neutralised, a situation which would be rare at best in normal, real speech.

The results of these experiments suggest that this may not be entirely the case for the modulation cue. Whilst the most heavily voiced unmasked stimuli (TNR = 10 dB) with no formant transition (PP = 0) are indeed judged voiced almost 100% of time regardless of whether they are modulated or not, at the other end of the transition continuum

	Transition Gap (PP)					
TNR	0	2	4	6	8	10
-50 dB	0.00	0.00	0.00	0.00	0.04	0.00
-10 dB	0.00	0.00	0.00	0.00	0.00	0.01
0 dB	0.02	0.03	0.00	0.00	0.00	0.00
10 dB	0.78	0.13	0.01	0.00	0.00	0.00

Table 5.2: *p*-values to 2 d.p. for Yates' χ^2 Test of Association. Voiced/voiceless response counts for modulated versus unmodulated stimuli are compared using a 2×2 contingency table analysis for combinations of TNR and Transition Gap. Unmasked stimuli.

(PP = 10), where the gap in formant transitions most strongly suggests a voiceless fricative (i.e., a stimulus which is clearly not neutralised), inclusion of modulation raises the voiced response by more than 20%, even in the most weekly voiced condition (TNR = -50 dB).

5.3.2 The Effect of Masking

A χ^2 significance test (modulated versus unmodulated) for all TNR and Transition Gap combinations (analogous to that performed for unmasked stimuli in Table 5.2 for masked stimuli) confirmed a significant difference at the $p < 0.05$ level in all cases.

In Figure 5.11, for the quietest voiced stimuli (TNR = -50 dB and -10 dB, top panels), the shape and level of the masked functions (thick lines) appear to follow that of the unmasked (thin lines) functions closely. Thus the vertical distance between modulated and unmodulated functions reflects that of the unmasked case, an approximate 30% increase in voiced responses.

For the stimuli with louder voicing (TNR = 0 dB and -10 dB) a different pattern is observed: for the masked stimuli, the Transition Gap effect appears to be neutralised. Both modulated and unmodulated functions for masked stimuli are flatter and lower than their unmasked counterparts, indicating that the Transition Gap cue is less salient for the masked stimuli as TNR rises. In the TNR = 10 dB case, the Transition Gap functions are almost flat and only the inclusion of AM increases voicing responses (from 31% to 58% at PP = 5).

In fact, for masked stimuli, the vertical distance between modulated and unmodulated functions remains constant throughout the transition gap range and does not narrow at shorter gaps as is the case for unmasked stimuli. Compare the modulated/unmodulated difference for stimuli with a PP = 2 (∼16 ms) Transition Gap: in the TNR = 0 dB case, the difference for unmasked stimuli is 9% compared to 50% for masked stimuli; for the TNR = 10 dB, the responses are 6% and 41% respectively.

For voiceless and quiet voicing stimuli, then, masking did not appear to have any effect on either subjects' overall voicing judgements, nor on their response to AM. For more strongly voiced stimuli, a different picture emerges: for unmasked stimuli, modulation was only effective in eliciting more voiced responses at long transitions; for masked stimuli, the length of the transition appears to have less (TNR = 0 dB) or close to no (TNR = 10 dB) effect and subjects appear to base their voicing judgements on the presence of AM.

Why should this be the case? Recall that the masker level was always 10 dB above that of the voicing; thus, as voicing level increased, so did the absolute level of the masker.

It is likely that at the loudest voicing level, the masker was loud enough to obscure the preceding vowel, thus impeding subjects' use of the vowel transition cue. In this case, it appears that subjects then relied primarily on AM as a cue to voicing.

5.3.3 The Effect of Modulation Depth

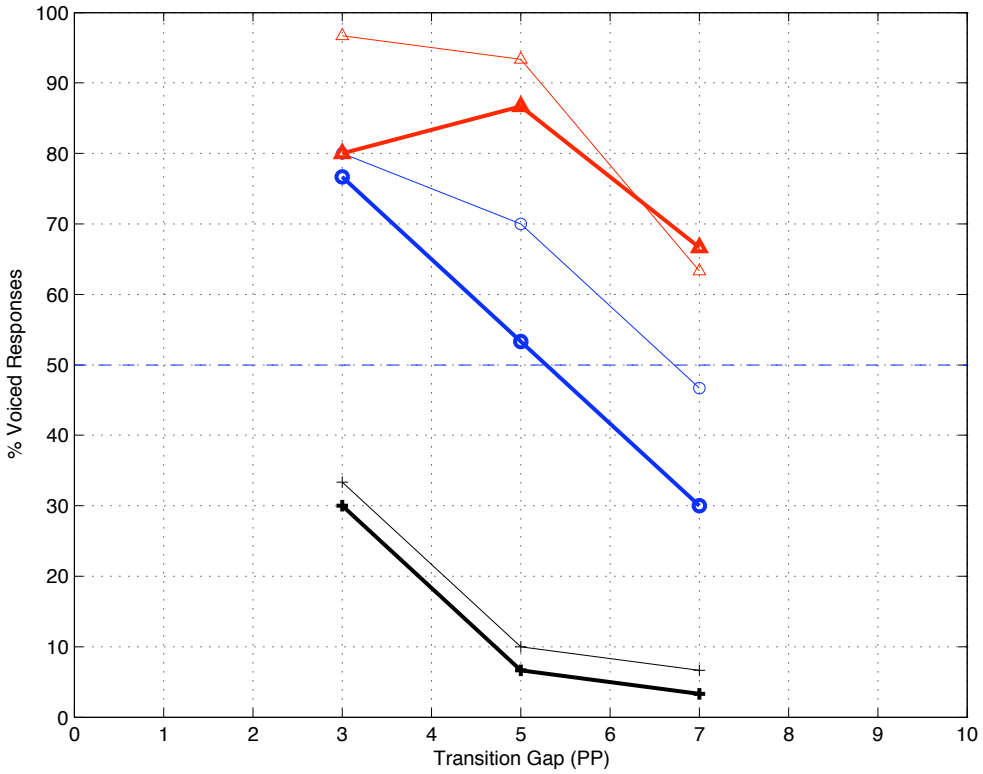

Figure 5.12: % 'voiced' responses at points along a formant transition continuum measured in pitch periods (PP) for three modulation conditions: $m = 0.1$ (black line, crosses), $m = 0.5$ (blue line, circles), $m = 0.7$ (red line, triangles) and two phase conditions, 0° (thin line) and 180° (thick line).

In an extension to the main experiment, modulation depth, m, and phase, ϕ, were introduced as variables whilst focusing only on the central part of the Transition continuum (3–7 pitch periods) where the function is steepest in the main experiment (see Figure 5.11). Results are presented in Figure 5.12.

Both m (thick lines; $m = 0.1$: black line, crosses; $m = 0.5$: blue line, circles; $m = 0.7$: red line, triangles) and ϕ (0°: thin line; 180°: thick line) have an effect on 'voiced' responses, which increase for larger modulation depths and the 0° phase condition. Thus, Transition trades not only with presence or absence of AM, but also with variables m and ϕ determining its depth and phase. For example, perceptual equivalence at 70% voiced responses is achieved by a 5 PP (~40 ms) transition, $m = 0.5$ and $\phi = 0°$ or by a 6.5 PP (~52 ms) transition, $m = 0.7$ and $\phi = 0°$ or 180°.

The figure suggests that the 0° condition (thin lines) produces marginally more voiced

responses than the 180° condition (thick lines) and that the effect is consistent across all values of m and PP with only a single exception. A χ^2 test (2×2 contingency table) confirms that the overall phase effect is significant at $p = 0.02$.

Post-hoc analysis, however, shows that whilst the phase effect is consistent, it is only significant at $p < 0.05$ for the $m = 0.7$, PP = 3 condition (evidenced by the wider gap between thin and thick red lines at PP = 3) and only significant at $p < 0.10$ for $m = 0.5$, PP = 5 and 7. So, whilst there is a small but significant and consistent effect for phase (limited to approximately 1 PP equivalence) in general, it is most prominent for the $m = 0.5$ condition.

The effect for AM depth is more pronounced. A 3×2 contingency table χ^2 analysis (m versus voiced/unvoiced responses) shows a significant effect for m at $p < 0.00$ and the difference in voiced responses for different m values at points along the formant transition continuum (vertical distance between functions) is significant in all but a single case (PP = 3, $\phi = 180°$, where $m = 0.5$ and 0.7 both produce approximately 80% voiced responses).

A point of interest arises as to how the results for $m = 0.1$ (black lines) relate to those of the main experiment, where voiced responses to the unmodulated stimuli under similar conditions (TNR = 0 dB, bottom-left panel: thin black line) were substantially higher. For example, in the extension, at PP = 3, 30% voiced responses are recorded; this contrasts with 70% in the main experiment for a *completely unmodulated stmulus*. Perceptual adaptation to the very deeply modulated $m = 0.7$ stimulus might explain this somewhat surprising result: assuming $m = 0.7$ is always detectible and perceptually salient, listeners may adapt and 'voiced' responses thus shift downwards in the $m = 0.1$ condition where modulation is probably undetectable in the majority of cases. The possibility of adaptation could be avoided by testing different AM depths during different sessions or on different subjects.

Nonetheless, results suggest that AM depth and phase relation to the voicing affect the weight accorded to modulation as a cue: deeper and in-phase modulation elicit more voiced responses. The jump in voicing responses when m is increased from 0.1 to 0.5 and correspondingly for 0° stimuli can almost certainly be attributed to increased detection of AM, since at $m = 0.1$, AM is probably not detected in the majority of cases. At $m = 0.5$, modulation should be well above the detection threshold considering the other parameters (TNR = 0 dB, VFR = 15 dB): in Experiment 5 of the psychoacoustic study, the detection threshold with VFR = 15 dB was found to be between -10 and -12 dB (0.25–0.32). Thus, the increased voiced responses obtained when raising m to 0.7 must be attributed to increased perceptual saliency, since it is assumed that AM is already detectible.

The conclusion, then, is that AM of frication noise is not necessarily a binary cue parameter and is perceptually integrated in a detailed and fine-grained manner along with a variety of other cues into a holistic representation of the voicing distinction. This study has shown how this is the case for formant transition structure, voicing amplitude and amplitude modulation of the frication noise component. There are, undoubtedly, numerous other acoustic characteristics that can be integrated into the perceptual representation of the voicing cue — some that have been mentioned in this thesis and others that are possibly unknown.

5.4 Summary

In a phonetic cue-trading experiment using engineered real speech, subjects' judgement of fricative voicing was measured as a function of the duration of a gap in the formant transition from the vowel leading into frication. A longer gap is characteristic of voiceless fricatives. Amplitude modulating the frication was found to induce an increase in voiced responses of between 13% (at PP = 0) and 30% (at PP = 6), with the perceptual equivalence of reducing the formant transition gap by approximately 2 pitch periods (∼16 ms). The effect of imposing AM was found to operate across levels of voicing amplitude and positions along the formant transition continuum with few exceptions, suggesting that AM is a robust cue and not only referred to when others are 'neutralised'. Furthermore, when loud masking obscures the voicing and formant transition cues, subjects are strongly biased to voiced responses by the presence of AM.

Results also suggest that increasing the perceptual saliency of AM (with deeper, in-phase modulation) increases its weight as a cue even when the existing modulation is over the estimated detection threshold. AM, then, appears to be integrated into the voicing cue percept in fricatives in a fine-grained manner that allows for its depth and phase to affect judgements of voicing.

Chapter 6

Conclusion

The objective of this study was to establish if amplitude modulation in the noise component of voiced fricatives sounds in speech could cue the phonological voicing distinction. In pursuing this objective, a bottom-up approach was adopted: research progressed through acoustic, then psychoacoustic and finally speech perception investigations.

The principal contribution of this study has been to introduce a new cue to the voicing distinction: amplitude modulation of the frication component. The perceptual relevance of AM has previously been hinted at, but listeners' use of the cue has not previously been confirmed in speech perception experiments.

The overall contribution of this study, however, is not limited to the principal research question. The multidisciplinary approach has produced relevant by-products in a number of separate fields.

6.1 Cues to the Voicing Distinction and Theories of Speech Perception

In the existing view of the perception of the voicing distinction in fricatives, three 'cues' have traditionally dominated research: the presence of amplitude of voicing, the duration of frication and in-vowel cues such as duration or formant transition characteristics (Section 2.3.3). There is, of course, debate as to which are primary or most perceptually salient.

In recent years, researchers have begun to accept the inadequacy of studying speech transmission and perception in 'perfect' laboratory conditions that do not replicate the noisy, reverberant and generally sub-optimal conditions of real conversations. Much of traditional perceptual theory, with its emphasis on discrete 'cues', may have thus

underestimated both the inherent difficulty of the speech transmission task and the importance of sub-optimal acoustic conditions.

Thus the evaluation of the principal objective of this research is one that applies not only to the specific 'cue' in question, but to all possible acoustic cues for all speech sounds. It has been established that subjects *can* use AM to distinguish voiced fricatives from their unvoiced counterparts, but the question remains as to whether they actually *do*.

There are limitations to the current work that prevent us from answering this question with certainty. The cue-trading experiment used limited parameters: only the fricative [z] was investigated and only in intervocalic position. Although results can presumably be generalised to other fricatives and positions, it is an assumption that requires further investigation in order to confirm. A more wide-ranging cue-trading experiment using all fricatives in all positions is ideally required in order to confirm the general availability of AM as a cue.

A further possibility would be to extend the work of Strope and Alwan (2001) with simulations on corpora of recorded fricatives using the knowledge obtained in the psychoacoustic study. Detectability and consequently availability as a cue could be determined based on the data from the psychoacoustic study: modulation frequency, modulation depth, phase relationship of modulator to the voicing, amplitude of voicing, segment duration, amplitude of vowel environment.

Still, the question of whether listeners actively use AM as a cue to voicing now falls into the general realm of current theoretical work on speech perception. This thesis adds to the mounting body of evidence in favour of a complex perceptual integration of all available phonetic information in the speech signal, including AM of frication. It is a question of current work in theories of speech perception to identify the underlying mechanisms responsible for this integration, whether it be reference to articulatory gestures, or some other framework.

6.2 Accuracy of Speech Synthesis

In improving the quality and intelligibility of synthetic speech, measurements of m made in the acoustic study could be reverse-engineered as control parameters for the latest articulatory-based speech synthesisers (e.g., Haskins CASY, SAPWindows for European Portuguese), or be used as direct input parameters to formant-type synthesisers [1]. Synthesisers of this type, based on fluid parameters that replicate the way in which sounds are articulated by speakers, may be the most promising route to truly

[1] Speech synthesis techniques based on concatenation of units of recorded speech such as diphones already encode AM in frication noise as it is present in the input signal

high quality synthesised speech, overcoming the glitches of the hitherto preferred concatenative methods, such as transitions between segments and the difficulty in full dynamic control of intonation and emotional characteristics (Whalen, 2003).

Fricatives are being successfully reproduced in modern articulatory synthesis systems but the level of detail needed to reproduce AM has not been reached (Stevens and Hanson, 2003; Teixeira et al., 2003). In fact, AM has so far been included in only rudimentary form in a number of older formant-type synthesis systems (Section 2.3.1). This could be partly because the perceptual role of AM was poorly understood, but is presumably largely attributable to a lack of accurate data that could control synthesiser parameters. Of particular relevance is the form of the functions relating modulation depth, m, to voicing strength, v, meaning that modulation of frication noise could be easily controlled by a synthesiser based on the latter. Fine-grained control could be achieved by applying the different functions that were measured for different fricative places of articulation or different speakers. Further research could build up a comprehensive library of m/v functions for many fricative places, contexts, speakers and perhaps even emotional conditions that could provide even greater control of AM in synthesisers.

A clearer direction for further work in this area, however, is the collection of phase data which was lacking from this study. The combination of modulation depth and phase provides a complete description of AM characteristics and would be required for fully accurate synthesis.

6.3 Models of AM Generation in the Vocal Tract

Although it is not yet clear which phenomenon causes AM of frication noise in the vocal tract, the results of the acoustic measurements showed unequivocally that m is related directly to voicing strength, a previously undemonstrated relationship in speech (Section 3.4.1). Furthermore, the nature of the relationship was shown to be non-linear, with saturation of m under the influence of strong voicing. As was pointed out, this saturation occurs well before $m = 1$, thus indicating that it is an effect of the physical modulation mechanism rather than a mathematical limitation.

Such results appear to lend support to the 'forcing wave' explanation of AM generation (Section 2.1.2) where the acoustic wave set up by voicing regularises turbulence production at the constriction. However, further theoretical work is needed to establish the predictions for the m/v relationship (as well as other phenomena observed in the acoustic study, such as higher m for [z]) made by other AM frameworks such as the static approach or the explanation invoking glottal vortices.

Regardless of the explanation for AM, further work in both physical and computational modelling is required to illuminate the generation mechanism before it can be explained with confidence.

Experiments using a physical replica of the vocal tract are currently being undertaken by Barney and Jackson. Using oscillating shutters to simulate glottal action, the research attempts to model aerodynamic conditions in the vocal tract that lead to AM of the frication noise. Preliminary measurements suggest that this is a promising line of enquiry (Barney and Jackson, 2006, 2007), with the particular benefit of being able to directly control parameters such as voicing strength (through force generated by the mechanical shutters) which can only be indirectly measured in human subjects. The primary need is for measurements of airflow conditions at a larger number of points superior to the glottis, continuing the work of Barney et al. (1999). Flow visualisation techniques might also be employed in establishing where modulation is created.

Further work in computational simulation of the aerodynamic and aeroacoustic conditions during frication could also shed light on the AM generation problem. The technique has already been applied to other areas of fricative study (Sinder, 1999; McGowan et al., 1995; R.S.McGowan and M.S.Howe, 2007) and could be adapted to the specific problem of voiced frication.

6.4 AM research

The results of psychoacoustic experiments 1–5 help to plug the gap in the literature regarding SAM detection in the presence of a tone whose frequency is equal to that of the modulator. Only a handful of studies have tackled this question directly, with the indication that the phase relationship between tone and modulator is crucial. Given that distortion tone products are restricted to very narrow noise bands (with bandwidth less than $4f_m$), the general conclusion is that an 'additive' or 'multiplicative' mechanism is responsible for the interation: residual energy from the low-frequency tone either adds to the output of auditory filters in the spectral area of the carrier or modifies its sensitivity. With the contribution of data from this study along with original data on 3 kHz bandwidth signals from Wakefield and Viemeister (1985), an interesting direction for future work arises.

The interaction effect between tone and carrier could depend on the bandwidth of the carrier. One hypothesis is that wider noise bands, with outputs over a wider range of auditory filters, could be less affected by the low-frequency signal. An AM detection experiment with interacting sinusoid could examine the effect of noise carrier

bandwidths ranging from very narrow to broadband, in order to establish whether this is the case.

Much emphasis is currently being placed on integrating the findings of research in AM detection and discrimination with traditional results from the 'audio' domain such as signal detection in noise and nonsimultaneous masking. The objective of such work is to unify the explanatory framework for both sets of results under a single theory or model. Recent advances in this field by Torsten Dau and coworkers has resulted in 'CASP', a computational model with a first-order 150 Hz modulation lowpass filter and modulation bandpass filtering at its core (Jepsen et al., 2008; Dau et al., 1997). The model is able to explain SAM detection and phenomena such as modulation masking with relative accuracy. Given that the objective of such auditory models is to explain a wide range of psychoacoustic phenomena, further work could compare the results presented in this thesis to simulations of processing of the CASP (or other) model on the types of signal used in the experiments. Future iterations of this and other such models should be able to explain the interaction of tone with AM detection and the importance of its amplitude and phase relationship to the modulating signal.

6.5 AM Detection in Fricative-Like Stimuli and the Bridge Between Psychoacoustics and Speech Perception

Voiced fricatives represent complex signals whose analysis from a psychoacoustic perspective is complicated by numerous factors. Research completed for these thesis has sought to systematically clarify how listeners hear AM of a noise carrier in the context of a simultaneous tone, whilst gradually introducing elements typical of VFs such as short noise durations, harmonic voicing instead of pure tone and a loud vowel environment.

Thus, this study has established how AM is detected in fricatives at a basic psychophysical level and how this perception relates to the phonological voicing classification. Little attempt has previously been made to relate elements of speech perception and classification to basic psychoacoustic processes. Even restricting ourselves to the case of voiced fricatives, many similar investigations suggest themeselves: what are the signal detection thresholds for voicing in fricative noise? Or vice versa? How do they vary for the spectral shaping of the noise (fricative place of articulation)? What is discrimination performance like for formant transitions considering characteristics of different fricatives etc? Of course, this line of enquiry extends beyond voiced fricatives to the entire collection of speech segments. In summary, speech research might benefit from asking how listeners hear cues before asking whether or how they *use* them.

The results of perceptual experiment 5 are particularly important to this line of re-

search. The literature lacks examples of experiments investigating basic psychacoustic processes such as signal detection, discrimination where the experiment is adapted or repeated using speech stimuli instead of combinations of sounds that are not heard as speech. Given the repeated suggestion of a 'speech mode' of perception, it was not unreasonable to assume that when listeners hear stimuli as speech, their performance on basic listening tasks might be affected. A few studies were identified that appeared to confirm this. The present research has presented a significant contribution to this research problem: it appears that AM detection is affected by listeners hearing stimuli as speech. The implications of this proposition for speech research are important. If detection and discrimination are impaired in the speech context, perhaps listeners are unable to 'hear' acoustic properties that have traditionally been thought of as cues. A concrete example: noise duration has traditionally been thought of as a cue to the voicing distinction in fricatives. But in the context of speech, are listeners able to discriminate durations to the accuity required to systematically categorise between voiced and voiceless? There appears to be the opportunity for a significant improvement in our understanding of the bridge between psychoacoustics and speech perception.

Appendix A

IPA Chart

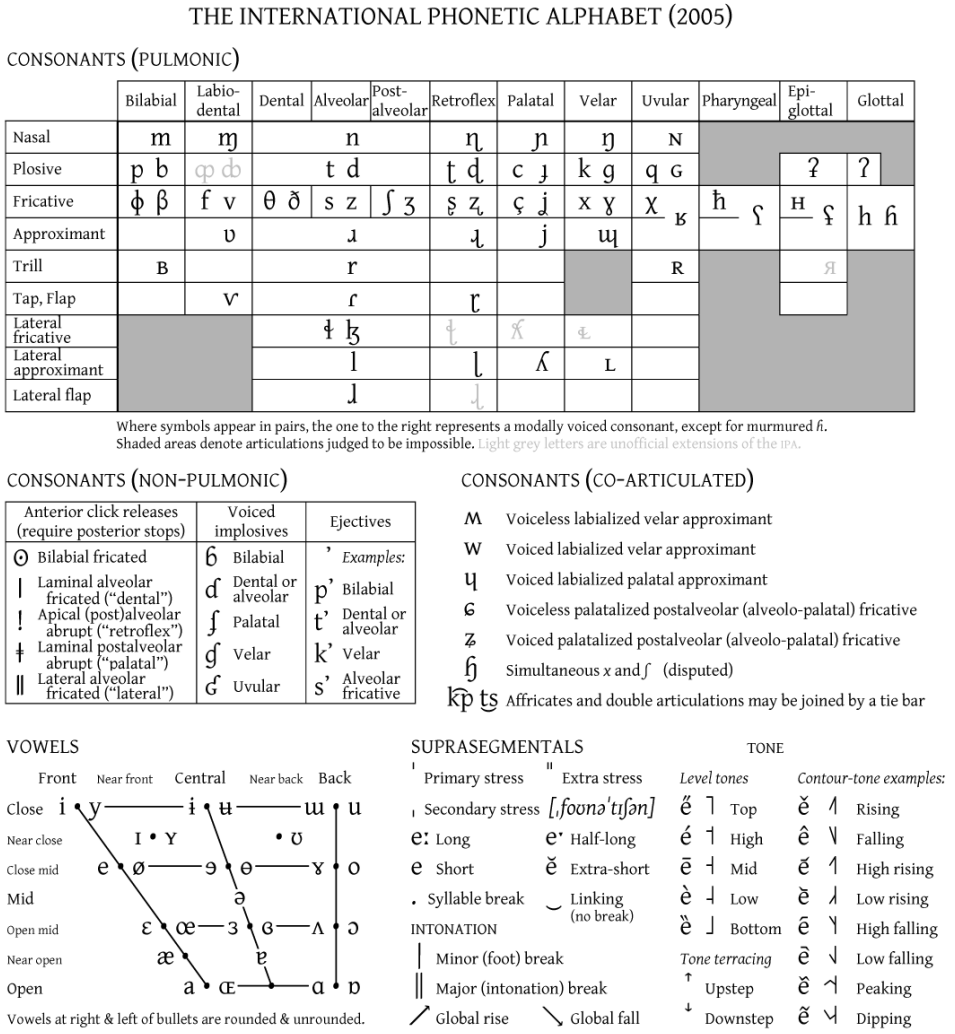
THE INTERNATIONAL PHONETIC ALPHABET (2005)
CONSONANTS (PULMONIC)
Bilabial
Labio-dental
Dental
Alveolar
Post-alveolar
Retroflex
Palatal
Velar
Uvular
Pharyngeal
Epi-glottal
Glottal
Nasal
Plosive
Fricative
Approximant
Trill
Tap, Flap
Lateral fricative
Lateral approximant
Lateral flap
Where symbols appear in pairs, the one to the right represents a modally voiced consonant, except for murmured ɦ.
Shaded areas denote articulations judged to be impossible. Light grey letters are unofficial extensions of the IPA.
CONSONANTS (NON-PULMONIC)
Anterior click releases (require posterior stops)
Bilabial fricated
Laminal alveolar fricated ("dental")
Apical (post)alveolar abrupt ("retroflex")
Laminal postalveolar abrupt ("palatal")
Lateral alveolar fricated ("lateral")
Voiced implosives
Bilabial
Dental or alveolar
Palatal
Velar
Uvular
Ejectives
Examples:
Bilabial
Dental or alveolar
Velar
Alveolar fricative
CONSONANTS (CO-ARTICULATED)
Voiceless labialized velar approximant
Voiced labialized velar approximant
Voiced labialized palatal approximant
Voiceless palatalized postalveolar (alveolo-palatal) fricative
Voiced palatalized postalveolar (alveolo-palatal) fricative
Simultaneous x and ʃ (disputed)
Affricates and double articulations may be joined by a tie bar
VOWELS
Front
Near front
Central
Near back
Back
Close
Near close
Close mid
Mid
Open mid
Near open
Open
Vowels at right & left of bullets are rounded & unrounded.
SUPRASEGMENTALS
Primary stress
Extra stress
Secondary stress
Long
Half-long
Short
Extra-short
Syllable break
Linking (no break)
INTONATION
Minor (foot) break
Major (intonation) break
Global rise
Global fall
TONE
Level tones
Top
High
Mid
Low
Bottom
Tone terracing
Upstep
Downstep
Contour-tone examples:
Rising
Falling
High rising
Low rising
High falling
Low falling
Peaking
Dipping
DIACRITICS
Diacritics may be placed above a symbol with a descender, as ŋ̊. Other IPA symbols may appear as diacritics to represent phonetic detail: tˢ (fricative release), bʱ (breathy voice), ʔa (glottal onset), ᵊ (epenthetic schwa), oᵅ (diphthongization).
SYLLABICITY & RELEASES
Syllable
Non-syllable
(Pre)aspirated
Nasal release
Lateral release
No audible release
PHONATION
Voiceless or Slack voice
Modal voice or Stiff voice
Breathy voice
Creaky voice
Strident
Linguolabial
PRIMARY ARTICULATION
Dental
Apical
Laminal
Advanced
Retracted
Centralized
SECONDARY ARTICULATION
Labialized
Palatalized
Velarized
Pharyngealized
Velarized or pharyngealized
Mid-centralized
More rounded
Less rounded
Nasalized
Rhoticity
Advanced tongue root
Retracted tongue root
Lowered (β is a bilabial approximant)
Raised (ɻ is a voiced alveolar non-sibilant fricative)

Figure A.1:

Appendix B

List of Randomised Sentences used in Recording Fluent Speech Fricative Corpus

What does eefer mean?
What does eether mean?
What does arjer mean?
What does arther mean?
What does arther mean?
What does eeser mean?
What does arther mean?
What does arther mean?
What does eefer mean?
What does oother mean?
What does oofer mean?
What does oother mean?
What does arjer mean?
What does eefer mean?
What does eesher mean?
What does eejer mean?
What does arjer mean?
What does arser mean?
What does eether mean?
What does oother mean?
What does arver mean?

What does arther mean?
What does arver mean?
What does ooser mean?
What does oother mean?
What does arjer mean?
What does oover mean?
What does oother mean?
What does eesher mean?
What does eether mean?
What does oother mean?
What does eeser mean?
What does oover mean?
What does arjer mean?
What does oosher mean?
What does arzer mean?
What does oozer mean?
What does oother mean?
What does eeser mean?
What does ooser mean?
What does eeser mean?
What does oother mean?

What does arsher mean?
What does oofer mean?
What does eeser mean?
What does oother mean?
What does oofer mean?
What does eever mean?
What does eejer mean?
What does oojer mean?
What does eether mean?
What does arfer mean?
What does arzer mean?
What does arfer mean?
What does arjer mean?
What does eesher mean?
What does arser mean?
What does oosher mean?
What does arfer mean?
What does arther mean?
What does arther mean?
What does arther mean?
What does arfer mean?

What does oother mean?
What does eeser mean?
What does arfer mean?
What does oosher mean?
What does arver mean?
What does ooser mean?
What does eever mean?
What does ooser mean?
What does oozer mean?
What does oover mean?
What does arther mean?
What does eever mean?
What does eesher mean?
What does eezer mean?
What does eever mean?
What does oover mean?
What does oother mean?
What does eejer mean?
What does eezer mean?
What does arser mean?
What does oozer mean?
What does arsher mean?
What does arzer mean?
What does oojer mean?
What does ooser mean?
What does eether mean?
What does arther mean?
What does eether mean?
What does arsher mean?
What does oosher mean?
What does oojer mean?
What does arser mean?
What does eether mean?
What does eever mean?
What does eezer mean?
What does eefer mean?
What does arther mean?
What does eever mean?

What does oojer mean?
What does arver mean?
What does oother mean?
What does oozer mean?
What does eeser mean?
What does arsher mean?
What does arzer mean?
What does oother mean?
What does eefer mean?
What does oozer mean?
What does eefer mean?
What does eejer mean?
What does arser mean?
What does oother mean?
What does eejer mean?
What does ooser mean?
What does oozer mean?
What does arsher mean?
What does eether mean?
What does arther mean?
What does eefer mean?
What does oofer mean?
What does arfer mean?
What does arser mean?
What does eever mean?
What does eesher mean?
What does eeser mean?
What does arther mean?
What does arsher mean?
What does arzer mean?
What does oozer mean?
What does arther mean?
What does eefer mean?
What does oosher mean?
What does eether mean?
What does oojer mean?
What does arsher mean?
What does oozer mean?
What does arzer mean?

What does eether mean?
What does oother mean?
What does arver mean?
What does oover mean?
What does arjer mean?
What does oozer mean?
What does oover mean?
What does arfer mean?
What does eezer mean?
What does arther mean?
What does arzer mean?
What does oover mean?
What does oofer mean?
What does oover mean?
What does arfer mean?
What does eesher mean?
What does eether mean?
What does eezer mean?
What does oofer mean?
What does eejer mean?
What does eezer mean?
What does arzer mean?
What does eever mean?
What does eefer mean?
What does eether mean?
What does arser mean?
What does eether mean?
What does arver mean?
What does eejer mean?
What does arsher mean?
What does arther mean?
What does ooser mean?
What does eesher mean?
What does eesher mean?
What does arver mean?
What does eejer mean?
What does eether mean?
What does eesher mean?
What does oosher mean?

What does eeser mean?
What does ooser mean?
What does arfer mean?
What does oofer mean?
What does arver mean?
What does oofer mean?
What does oosher mean?
What does oojer mean?
What does oofer mean?
What does eether mean?
What does arver mean?
What does oover mean?
What does ooser mean?
What does oojer mean?
What does eezer mean?
What does arther mean?
What does arsher mean?
What does eether mean?
What does oother mean?
What does oother mean?
What does arther mean?
What does eejer mean?
What does oosher mean?
What does oojer mean?
What does arjer mean?
What does eever mean?
What does eether mean?
What does arjer mean?
What does oosher mean?
What does arser mean?
What does eether mean?
What does eezer mean?
What does oojer mean?
What does eezer mean?
What does oother mean?
What does arser mean?

Appendix C

Definition of Statistical Tests

C.1 ANOVA

C.2 Yates' χ^2 Test

Yates' χ^2 Test differs from Pearson's test in that 0.5 is subtracted from the magnitude of the difference between observed and expected frequencies before squaring:

$$\chi^2 = \sum_{i=1}^{N} \frac{(|O_i - E_i| - 0.5)^2}{E_i}, \tag{C.1}$$

where O_i is an observed frequency, E_i is an expected or theoretical frequency and N is the number of cells in the contingency table.

Bibliography

Arkebauer, H., T. J. Hixon, and J. Hardy (1967). Peak intraoral air pressures during speech. *Journal of Speech and Hearing Research 10*, 196–208.

Bacon, S. and N. Viemeister (1985). Temporal modulation transfer functions in normal-hearing and hearing-impaired subjects. *Audiology 24*, 117–134.

Bacon, S. P. and J. Lee (1997 Jun). The modulated-unmodulated difference: effects of signal frequency and masker modulation depth. *J Acoust Soc Am 101*(6), 3617–3624.

Bacon, S. P., J. Lee, D. N. Peterson, and D. Rainey (1997 Mar). Masking by modulated and unmodulated noise: effects of bandwidth, modulation rate, signal frequency, and masker level. *J Acoust Soc Am 101*(3), 1600–1610.

Badin, P. and G. Fant (1989, September). Fricative production modelling: Aerodynamic and acoustic data. In *Proceedings of the European Conference on Speech Communication and Technology, Paris*.

Bailey, P. J. and Q. Summerfield (1980). Information in speech: observations on the perception of [s]-stop clusters. *J Exp Psychol Hum Percept Perform 6*(3), 536–563.

Barney, A. and P. J. B. Jackson (2006). Modulation of frication noise in a dynamic mechanical model of the larynx and vocal tract. In *J. Acoust. Soc. Am., 119 (5, Pt. 2):*.

Barney, A. and P. J. B. Jackson (2007). Aerodynamically-based parametric description of the noise envelope in voiced fricatives. In *J. Acoust. Soc. Am., 121 (5, Pt. 2): 3122 A, Salt Lake City, UT,*.

Barney, A., C. H. Shadle, and P. Davies (1999). Fluid flow in a dynamic mechanical model of the vocal folds and tract. 1. measurements and theory. *J. Acoust. Soc. Am. 105*(1), 444–455.

Baum, S. R. and S. E. Blumstein (1987, September). Preliminary observations on the use of duration as a cue to syllable-initial fricative consonant voicing in English. *J. Acoust. Soc. Am. 82*(3), 1073–1077.

Beautemps, D., P. Badin, and R. Laboissière (1995). Deriving vocal-tract area functions from the midsagittal profiles and formant frequencies: A model for vowels and fricative consonants based on experimental data. *Speech Communication 16*, 28–47.

Blumstein, S. E. and K. N. Stevens (1979). Acoustic invariance in speech production: evidence from measurements of the spectral characteristics of stop consonants. *J Acoust Soc Am 66*(4), 1001–1017.

Blumstein, S. E. and K. N. Stevens (1980). Perceptual invariance and onset spectra for stop consonants in different vowel environments. *J Acoust Soc Am 67*(2), 648–662.

Blumstein, S. E. and K. N. Stevens (1981). Phonetic features and acoustic invariance in speech. *Cognition 10*(1-3), 25–32.

Blumstein, S. E., K. N. Stevens, and G. N. Nigro (1977). Property detectors for bursts and transitions in speech perception. *J Acoust Soc Am 61*(5), 1301–1313.

Borzone de Manrique, A. M. and M. I. Massone (1981, April). Acoustic analysis and perception of Spanish fricative consonants. *J. Acoust. Soc. Am. 69*(4), 1145–1153.

Bregman, A. (1990). *Auditory Scene Analysis: The Perceptual Organisation of Sound*. MIT, Cambridge, MA.

Burns, E. and N. Viemeister (1976). Nonspectral pitch. *J. Acoust. Soc. Am. 60*, 863–868.

Burns, E. and N. Viemeister (1981). Played again sam: Further observation on the pitch of amplitude modulated noise. *J. Acoust. Soc. Am. 70*, 1655–1660.

Camargo, A., F. Azuaje, H. Wang, and H. Zheng (2008). Permutation-based statistical tests for multiple hypotheses. *Source Code for Biology and Medicine 3*(15), 3–15.

Coker, C. H., M. H. Krane, B. Y. Reis, and R. A. Kubli (1996). Search for unexplored effects in speech production. In *Proc. Int. Conf. Spoken Language Processing 1996,* Philadelphia, PA, Volume 14(6), pp. 415–422.

Cole, R. A. and W. E. Cooper (1975, December). Perception of voicing in English affricates and fricatives. *J. Acoust. Soc. Am. 58*(6), 1280–1287.

Cosgrove, P., J. Wilson, and R. Patterson (1989). Formant transition detection in isolated vowels with transitions in initial and final position formant transition detection in isolated vowels with transitions in initial and final position. In *Acoustics, Speech, and Signal Processing, 1989. ICASSP-89., 1989*, Volume 1, pp. 278 – 281. IEEE.

Cox, T. (2008). Metrics: Roughness and fluctuation strength.

Crow, S. and F. Champagne (1971). Orderly structure in jet turbulence. *Journal of Fluid Mechanics 48*(3), 547–591.

Crystal, T. H. and A. S. House (1982, September). Segmental durations in connected speech signals: Preliminary results. *J. Acoust. Soc. Am. 72*(3), 705–716.

Crystal, T. H. and A. S. House (1988, April). Segmental durations in connected speech signals: Current results. *J. Acoust. Soc. Am. 83*(4), 1553–1573.

Daniel, P. and R.Weber (1997). Psychoacoustical roughness: Implementation of an optimized model. *Acustica 83*, 113–123.

Dau, T., B. Kollmeier, and A. Kohlrausch (1997). Modelling auditory processing of amplitude modulation. 1. detection and masking with narrow-band carriers. *J. Acoust. Soc. Am. 102*(5), 2892–2905.

Davies, P. O. A. L. (1981). Flow-acoustic coupling in ducts. *Journal of Sound and Vibration 77*(2), 191–209.

Denes, P. (1955, July). Effect of duration on the perception of voicing. *J. Acoust. Soc. Am. 27*(4), 761–764.

Diehl, R. L. and K. R. Kluender (1990). On the categorization of speech sounds. In S. R. Harnad (Ed.), *Categorical Perception*. Cambridge University Press.

Diehl, R. L., A. J. Lotto, and L. L. Holt (2004). Speech perception. *Annu Rev Psychol 55*, 149–179.

Dubrovskii, N. and L. Tumarkina (1967). Investigation of the human perception of amplitude-modulated noise. *Soviet Physics and Acoustics-USSR 13*, 41–47.

Eddins, D. (1993). Amplitude modulation detection of narrow-band noise: Effects of absolute bandwidth and frequency region. *J. Acoust. Soc. Am. 93*(1), 470–479.

Ewert, S. D. and T. Dau (2004). External and internal limitation in amplitude-modulation processing. *J. Acoust. Soc. Am. 116*(1), 478–490.

Fant, G. (1960). *Acoustic Theory of Speech Production*. The Hague, Netherlands: Mouton.

Fastl, H. (1976). Temporal masking effects: 1. broad band noise masker. *Acustica 35*, 287–302.

Flanagan, J. L. and L. Cherry (1969). Excitation of vocal-tract synthesizers. *J. Acoust. Soc. Am. 45*(3), 764–769.

Flanagan, J. L. and M. G. Saslow (1957). Pitch discrimination for synthetic vowels. *J. Acoust. Soc. Am. 30*(5), 435–442.

Fletcher, H. (1940). Auditory patterns. *Rev. Mod. Phys. 12*, 47–65.

Formby, C. and K. Muir (1988). Modulation and gap detection for broadband and filtered noise signals. *J. Acoust. Soc. Am. 84*(2), 545–550.

Fowler, C. A. (1991). Auditory perception is not special: we see the world, we feel the world, we hear the world. *J Acoust Soc Am 89*(6), 2910–2915.

Fowler, C. A. (1996). Listeners do hear sounds, not tongues. *J Acoust Soc Am 99*(3), 1730–1741.

Grimault, N., S. P. Bacon, and C. Micheyl (2002 Mar). Auditory stream segregation on the basis of amplitude-modulation rate. *J Acoust Soc Am 111*(3), 1340–1348.

Grose, J. H. and J. W. r. Hall (1992). Comodulation masking release for speech stimuli. *J Acoust Soc Am 91*(2), 1042–1050.

Haggard, M. (1978). The devoicing of voiced fricatives. *Journal of Phonetics 6*, 95–102.

Hall, J. W., M. P. Haggard, and M. A. Fernandes (1984, Jul). Detection in noise by spectro-temporal pattern analysis. *J Acoust Soc Am 76*(1), 50–56.

Harris, G. (1963). Periodicity perception by using gated noise. *J. Acoust. Soc. Am. 35*, 1229–1233.

Healy, E. W. and S. P. Bacon (2006). Measuring the critical band for speech. *J Acoust Soc Am 119*(2), 1083–1091.

Heid, S. and S. Hawkins (1999). Synthesizing systematic variation at the boundaries between vowels and obstruents. In *Proc. ICPhs,* San Fransisco, pp. 511–514.

Hermes, D. J. (1991). Synthesis of breathy vowels: some research methods. *Speech Comm. 10*(5-6), 497–502.

Hixon, T. J. (1966). Turbulent noise sources for speech. *Folia Phoniatrica 18*, 168–182.

International Phonetic Association (1999). *Handbook of the International Phonetic Association*. Trumpington St., Cambridge, UK: Cambridge University Press.

Isshiki, N. and R. Ringel (1964). Airflow during the production of selected consonants. *Journal of Speech and Hearing Research 7*, 233–244.

Jackson, P. J. B. (2000). *Characterisation of Plosive, Fricative and Aspiration Components in Speech Production.* Ph. D. thesis, Dept. Electronics and Computer Science, University of Southampton.

Jackson, P. J. B. and C. H. Shadle (2000). Frication noise modulated by voicing, as revealed by pitch-scaled decomposition. *J. Acoust. Soc. Am. 108*(4), 1421–1434.

Jackson, P. J. B. and C. H. Shadle (2001). Pitch-scaled estimation of simultaneous voiced and turbulence-noise components in speech. *IEEE Trans. on Speech & Audio Proc. 9*(7), 713–726.

Jeong, H. (1999). *Sound Quality Analysis of Nonstationary Acoustic Signals.* Ph. D. thesis, Department of Mechanical Engineering, Korea Advanced Institute of Science and Technology (KAIST).

Jepsen, M. L., S. D. Ewert, and T. Dau (2008). A computational model of human auditory signal processing and perception. *J Acoust Soc Am 124*(1), 422–438.

Jongman, A. (1989, April). Duration of fricative noise required for identification of English fricatives. *J. Acoust. Soc. Am. 85*(4), 1718–1725.

Jongman, A., R. Wayland, and S. Wong (2000, September). Acoustic characteristics of English fricatives. *J. Acoust. Soc. Am. 108*(3), 1252–1263.

Klatt, D. H. (1980). Software for a cascade/parallel formant synthesizer. *J. Acoust. Soc. Am. 67*(3), 971–995.

Ladefoged, P. (2008). Representing linguistic phonetic structure. Unfinished draft available at http://www.linguistics.ucla.edu/people/ladefoge/.

Laver, J. (1994). *Principles of Phonetics.* Cambridge University Press.

Lee, J. and S. P. Bacon (1997). Amplitude modulation depth discrimination of a sinusoidal carrier: Effect of stimulus duration. *J. Acoust. Soc. Am. 101*(6), 3688–3693.

Levitt, H. (1970). Transformed up-down methods in psychoacoustics. *J. Acoust. Soc. Am. 49*(2), 467–477.

Lewis, D. E. and T. D. Carrell (2007). The effect of amplitude modulation on intelligibility of time-varying sinusoidal speech in children and adults. *Percept Psychophys 69*(7), 1140–1151.

Liberman, A., F. Cooper, D. Shankweiler, and M.Studdert-Kennedy (1967, November). Perception of the speech code. *Psychological Review 74*(6), 431–461.

Liberman, A. M. (1996). *Speech: A Special Code*, Chapter Introduction: Some Assumptions about Speech and How They Changed, pp. 1–46. Cambridge University Press.

Lighthill, M. (1952). On sound generated aerodynamically. I. General theory. In *Proceedings of the Royal Society*, Volume 211, pp. 564–587.

Lighthill, M. (1954). On sound generated aerodynamically.II. Turbulence as a source of sound. In *Proceedings of the Royal Society*, Volume 222, pp. 1–34.

Lofqvist, A., T. Baer, N.S.McGarr, and R.Seider-Story (1989). The cricothyroid muscle in voicing control. *J. Acoust. Soc. Am. 85*(3), 1314–1321.

Lofqvist, A., L. L. Koenig, and R. S. McGowan (1995). Vocal tract aerodynamics in /aCa/ utterances: Measurements. *Speech Comm. 16*, 50–66.

Mair, S. J. and C. H. Shadle (1996). The voiced/voiceless distinction in fricatives: EPG, acoustic and aerodynamic data. *Proceedings of the Institute of Acoustics 18*(9), 163–169.

Malécot, A. (1955). An experimental study of force of articulation. *Studia Linguistics 9*, 35–44.

Mann, V. A. and B. H. Repp (1980). Influence of vocalic context on perception of the [/textesh]-[s] distinction. *Perception and Psychophysics 28*, 213–228.

Massaro, D. W. and M. M. Cohen (1976, September). The contribution of fundamental frequency and voice onset time to the /zi/–/si/ distinction. *J. Acoust. Soc. Am. 60*(3), 704–717.

Massey and J. Smith (1998). *Mechanics of Fluids, 7th Edition.* Stanley Thornes Publishers.

McFadden, D. (1975 Apr). Beat-like interaction between periodic wave forms. *J Acoust Soc Am 57*(4), 983.

McFadden, D. (1988). Failure of a missing-fundamental complex to interact with masked and unmasked pure tones at its fundamental frequency. *Hearing Research 32*(1), 23–39.

McGowan, R. S., L. L. Koenig, and A. Lofqvist (1995). Vocal tract aerodynamics in /aCa/ utterances: Simulations. *Speech Communication 16*, 67–88.

Miller, G. A. and W. Taylor (1948). The perception of repeated bursts of noise. *J. Acoust. Soc. Am. 20*, 171–182.

Milosevic, M. A., A. M. Mitic, and M. S. Milosovic (2004). Parameters influencing noise estimation. *Facta Universitatis 2*(4), 277–284.

Moore, B. C. and B. R. Glasberg (1983). Growth of forward masking for sinusoidal and noise maskers as a function of signal delay; implications for suppression in noise. *J Acoust Soc Am 73*(4), 1249–1259.

Munson, Young, and Okiishi (1990). *Fundamentals of Fluid Mechanics*. Wiley Publishers.

Narayanan, S. S. (1995). *Fricative Consonants: An articulatory, acoustic and systems study*. Ph. D. thesis, Department of Electrical Engineering, University of California, Los Angeles, CA.

Narayanan, S. S. and A. A. Alwan (2000). Noise source models for fricative consonants. *IEEE Trans. on Speech & Audio Proc. 8*(2), 328–344.

Narayanan, S. S., A. A. Alwan, and K. Haker (1995). An articulatory study of fricative consonants using magnetic resonance imaging. *J. Acoust. Soc. Am. 98*(3), 1325–1347.

Oxenham, A. J. and B. C. Moore (1994). Modeling the additivity of nonsimultaneous masking. *Hear Res 80*(1), 105–118.

Pastel, L. (1987). Turbulent noise sources in vocal tract models. Master's thesis, MIT, Cambridge, MA.

Pickett, J. M. (1999). *The Acoustics of Speech Communication: Fundamentals, Speech Perception Theory and Technology*. Allyn and Bacon.

Pincas, J. (2004). The interaction of voicing and frication sources in speech: An acoustic study. Master's thesis, School of Electronics and Physical Sciences, University of Surrey.

Pincas, J. and P. J. Jackson (2004, June). Acoustic correlates of voicing-frication interaction in fricatives. In *Proceedings of the 'From Sound to Sense' Conference, MIT, Cambridge, MA, USA*.

Pirello, K., S. E. Blumstein, and K. Kurowski (1997, June). The characteristics of voicing in syllable-initial fricatives in American English. *J. Acoust. Soc. Am. 101*(6), 3754–3765.

Pohlmann, K. C. (2005). *Principles of Digital Audio*. McGraw-Hill Professional.

Pollack, I. (1969 Jan). Periodicity pitch for interrupted white noise–fact or artifact? *J Acoust Soc Am 45*(1), 237–238.

Remez, R. E., P. E. Rubin, D. B. Pisoni, and T. D. Carrell (1981). Speech perception without traditional speech cues. *Science 212*(4497), 947–949.

Repp, B. H. (1982). Phonetic trading relations and context effects: new experimental evidence for a speech mode of perception. *Psychol Bull 92*(1), 81–110.

Repp, B. H. and A. M. Liberman (1990). Phonetic category boundaries are flexible. In S. R. Harnad (Ed.), *Categorical Perception*. Cambridge University Press.

Repp, B. H., A. M. Liberman, T. Eccardt, and D. Pesetsky (1978). Perceptual integration of acoustic cues for stop, fricative, and affricate manner. *J Exp Psychol Hum Percept Perform 4*(4), 621–637.

Rodenburg, M. (1972). *Sensitivity of the Auditory System to Differences in Intensity*. Ph. D. thesis, Medical Faculty of Rotterdam.

Rodenburg, M. (1977). Investigation of temporal effects with amplitude modulated signals. In E.F.Evans and J.P.Wilson (Eds.), *Psychophysics and Physiology of Hearing*, pp. 429–437. London: Academic.

Rosen, S., A. Faulkner, and L. Wilson (1999). Adaptation by normal listeners to upward spectral shifts of speech: Implications for cochlear implants. *J. Acoust. Soc. Am. 106*(6), 3629–3636.

R.S.McGowan and M.S.Howe (2007). Compact green's functions extend the acoustic theory of speech production. *Journal of Phonetics 35*, 259–270.

Scully, C. (1990). Articulatory synthesis. In W. J. Hardcastle and A. Marchal (Eds.), *Speech Production and Speech Modelling*, pp. 151–186. Dordrecht, Netherlands: Kluwer Academic.

Scully, C., E. Castelli, E. Brearley, and M. Shirt (1992). Analysis and simulation of a speaker's aerodynamic and acoustic patterns for fricatives. *J. Phon. 20*, 39–51.

Shadle, C. (1990). Articulatory-acoustic relationships in fricative consonants in speech production and speech modelling. In W.J.Hardcastle and A. Marchal (Eds.), *Speech Production and Speech Modelling*, pp. 187–209. Kluwer Academic Publishers.

Shadle, C. H. (1985, March). The acoustics of fricative consonants. Technical Report 506, RLE, Massachusetts Institute of Technology.

Shadle, C. H. (1995). Modelling the noise source in voiced fricatives. In *Proc. 15th Int. Congress on Acoustics* Trondheim, Norway, Volume 3.

Shannon, R. V., F.-G. Zen, V. Kamath, J. Wygonski, and M. Ekelid (1995). Speech recognition with temporal cues. *Science 270*, 303–304.

Sheft, S. and W. A. Yost (1990). Temporal integration in amplitude modulation detection. *J Acoust Soc Am 88*(2), 796–805.

Simcox, C. and R. Hoglund (1971, March). Acoustic interactions with turbulent jets. In *Transactions of the ASME*, pp. 42–46.

Sinder, D. J. (1999). *Speech synthesis using an aeroacoustic fricative model*. Ph. D. thesis, Dept. Electrical Engineering, Rutgers University, New Brunswick, NJ.

Slaney, M. (1993). An efficient implementation of the patterson-holdsworth auditory filter bank. *Apple Computer Technical Report 35*, 1–50.

Smith, C. L. (1995). Contextual influences on devoicing of /z/ in American English. In *Proceedings of the 13th International Congress of the Phonetic Sciences*.

Smith, C. L. (1997). The devoicing of /z/ in American English: Effects of local and prosodic context. *Journal of Phonetics 25*, 471–500.

Soli, S. D. (1982, August). Structure and duration of vowels together specify fricative voicing. *J. Acoust. Soc. Am. 72*(2), 366–378.

Sondhi, M. M. and J. Schroeter (1987). A hybrid time-frequency domain articulatory speech synthesiser. *IEEE Trans. on Acoust., Speech & Sig. Proc. 35*(7), 955–967.

Sporer, T. and H. Schroder (1992, Dec.). Measuring tone masking noise. *J. Audio Eng. Soc. (Abstracts) 40*, 1038.

Stein, A., S. D. Ewert, and L. Wiegrebe (2005a, Oct). Perceptual interaction between carrier periodicity and amplitude modulation in broadband stimuli: a comparison of the autocorrelation and modulation-filterbank model. *J Acoust Soc Am 118*(4), 2470–2481.

Stein, A., S. D. Ewert, and L. Wiegrebe (2005b). Perceptual interaction between carrier periodicity and aplitude modulation in boradband stimuli: A comparison of the autocorrelation and modulation-filterbank model. *J. Acoust. Soc. Am. 118*, 2470–2481.

Stevens, K. and H. Hanson (2003, August 3-9). Production of consonants with a quasi-articulatory synthesizer. In *Proceedings of the 15th International Congress of Phonetic Sciences*, pp. 199–202.

Stevens, K. N. (1971). Airflow and turbulence noise for fricatives and stop consonants: Static considerations. *J. Acoust. Soc. Am. 50*(4), 1180–1192.

Stevens, K. N. (1998). *Acoustic Phonetics*. Cambridge, MA 02142-1493, USA: The MIT Press.

Stevens, K. N. and S. E. Blumstein (1978). Invariant cues for place of articulation in stop consonants. *J Acoust Soc Am 64*(5), 1358–1368.

Stevens, K. N., S. E. Blumstein, L. Glicksman, M. Burton, and K. Kurowski. (1992, May). Acoustic and perceptual characteristics of voicing in fricatives and fricative clusters. *J. Acoust. Soc. Am. 91*(5), 2979–3000.

Strickland, E. A. and N. F. Viemeister (1997). The effects of frequency region and bandwidth on the temporal modulation transfer function. *J. Acoust. Soc. Am. 102*(3), 1799–1810.

Strope, B. and A. A. Alwan (1998, June). Amplitude modulation cues for perceptual voicing distinctions in noise. In *Proceedings of 135th Meeting of the Acoustical Society of America*, Volume 103 (5), Part 2, pp. 2771.

Strope, B. P. and A. A. Alwan (2001). Modeling the perception of pitch-rate amplitude modulation in noise. In S.Greenberg and M.Slaney (Eds.), *Computational Models of Auditory Function*, pp. 315–327. IOS Press.

Tchorz, J. and B. Kollmeier (2002). Estimation of the signal-to-noise ratio with amplitude modulation spectrograms. *Speech Comm. 38*, 1–17.

Teixeira, A., L. M. T. Jesus, and R. Martinez (2003). Adding fricatives to the Portuguese articulatory synthesiser. In *EUROSPEECH*.

Teixeira, A. J. S., R. Martinez, L. M. T. Jesus, J. C. Principe, and F. A. C. Vaz (2005). Simulation of human speech production applied to the study and synthesis of european portuguese. *EURASIP Journal on Applied Signal Processing 9*, 1435–1448.

Terhardt, E. (1974). On the perception of periodic sound fluctuations. *Acustica 30*, 201–213.

Tritton, D. (1988). *Physical Fluid Mechanics, 2nd Edition*. Oxford University Press.

Verhey, J. L., D. Pressnitzer, and I. M. Winter (2003). The psychophysics and physiology of comodulation masking release. *Exp Brain Res 153*(4), 405–417.

Viemeister, N. (1973). Temporal modulation transfer functions for audition. *J. Acoust. Soc. Am. 53*, 312(A).

Viemeister, N. (1977). Temporal factors in audition: A systems analysis approach. In E.F.Evans and J.P.Wilson (Eds.), *Psychophysics and Physiology of Hearing*, pp. 419–428. London: Academic.

Viemeister, N. (1979). Temporal modulation transfer functions based upon modulation thresholds. *J. Acoust. Soc. Am. 66*(5), 1364–1380.

Viemeister, N. F. (1970). *Auditory Discrimination of Intensity, Internal Noise, and Temporal Processing*. Ph. D. thesis, Indiana University, Bloomington, IN.

Wakefield, G. H. and N. F. Viemeister (1985). Temporal interactions between pure tones and amplitude modulated noise. *J. Acoust. Soc. Am. 77*(4), 1535–1542.

Wakefield, G. H. and N. F. Viemeister (1990). Discrimination of modulation depth of sinusoidal amplitude modulation (SAM) noise. *J. Acoust. Soc. Am. 88*(1), 1367–1373.

Whalen, D. H. (2003, August). Articulatory synthesis: Advances and prospects. In *Proceedings of the 15th International Congress of the Phonetic Sciences, Barcelona*, pp. 175–177.

Wiegrebe, L. and R. D. Patterson (1999). Quantifying the distortion products generated by amplitude-modulated noise. *J Acoust Soc Am 106*(5), 2709–2718.

Yates, F. (1934). Contingency tables involving small numbers and the 2 test. *Journal of the Royal Statistical Society 1*, 217–235.

Yegnanarayana, B., C. d'Alessandro, and V. Darsinos (1998, January). An iterative algorithm for decomposition of speech signals into periodic and aperiodic components. *IEEE Trans. on Speech & Audio Proc. 6*(1), 11 pp.

Zhao, W., H.Frankel, and L.Mongeau (2000). Effects of trailing jet instability on vortex ring formation. *Phys. Fluids 12*(3), 589–596.

Zwicker, E. (1976). A model for predicting masking-period patterns. *Biol. Cybern. 23*, 49–60.

Zwicker, E. (1984). Dependence of post-masking on masker duration and its relation to temporal effects in loudness. *J Acoust Soc Am 75*(1), 219–223.

Zwicker, E. and H. Fastl (1999). *Psychoacoustics: Facts and Models, 2^{nd} Edition*. Springer-Verlag, Berlin.

Zwicker, E. and R. Feldtkeller (1967). *Das ohr als Nachrichtenemfanger*. Stuttgart: Hirzel-Verlag.